UNIVERZITA KARLOVA
PŘÍRODOVĚDECKÁ FAKULTA

Studijní program: Bioinformatika
Studijní obor: Bioinformatika

Anotace sekundární struktury proteinů
Protein secondary structure assignment

Jan Keil

BAKALÁŘSKÁ PRÁCE

Školitel: Mgr. Marian Novotný, Ph.D.

Praha 2018

Děkuji svému školiteli Mgr. Marianu Novotnému, Ph.D. za věnovaný čas, odborné konzultace a trpělivost.

Prohlašuji, že jsem tuto bakalářskou práci vypracoval sám pouze z citovaných zdrojů a literatury a na základě konzultací se svým školitelem. Dále prohlašuji, že jsem nepředložil práci ani její podstatnou část k získání jiného nebo obdobného akademického titulu.

Praha, 2018

Obsah

Abstrakt

Jako sekundární struktury proteinů jsou označovány konzervované motivy struktur vznikající díky slabým vazebným interakcím.

V úvodu práce se zaměřuji na popis a charakterizaci sekundárních struktur. Speciální pozornost je věnována popisu vzácnějších typů helikálních sekundárních struktur, jako jsou 3_{10}-helixy či π-helixy. Jádrem práce je však přehled a zhodnocení existujících metod, které umí ve 3D strukturách proteinů takovéto motivy identifikovat.

Klíčová slova

Helix, list, assignment, anotace, sekundární struktura, algoritmus.

Abstract

Protein secondary structures are conserved motifs of structures resulting from weak binding interactions.

At the beginning of this thesis I focus on description and characterization of secondary structures. Special attention is paid to the description of rarer types of helical secondary structures such as 3_{10}-helixes or π-helixes. At the core of the thesis, however, is an overview and an evaluation of existing methods that are able to identify such motifs in protein 3D structures.

Keywords

Helix, sheet, assignment, secondary structure, algorithm.

Seznam použitých zkratek:

3D	=	Třídimenzionální
AK	=	Aminokyselina
ASS	=	Secondary structure assignment (anotace sekundární struktury)
CM	=	Character Matrix (matice znaků)
DM	=	Distance Matrix (vzdálenostní matice)
HB	=	Hydrogen bond (vodíková vazba, vodíkový můstek)
MD	=	Molekulární dynamika
NMR	=	Nukleární magnetická rezonance
PDB	=	Protein Data Bank
RMS	=	Root Mean Square (efektivní hodnota)
SSE	=	Secondary Structure Elelments (elementy sekundárních struktur)
WoS	=	Web of Science

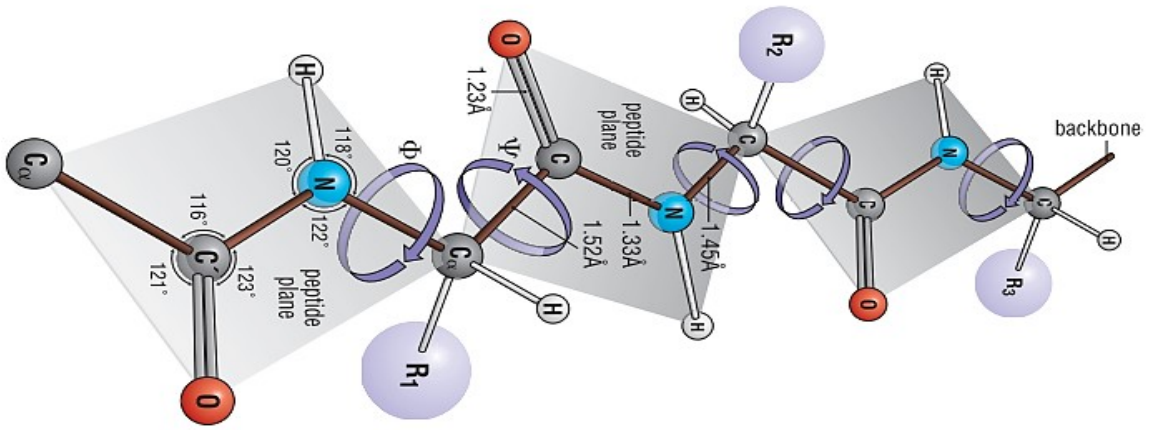
1 Úvod

Proteiny představují jeden ze základních stavebních kamenů každého živého organismu a podílejí se téměř na každém ději, který je s jeho existencí spjat. Pro pochopení průběhu složitých biochemických procesů v živých organismech je nutné zjistit, jak v něm jednotlivé proteiny fungují, na jakých procesech se podílejí. Klíčový význam pro zjištění funkce proteinů má právě určování jejich struktury. Znalost struktury proteinu je nezbytná k odhalení vztahu mezi genem a jeho projevem – funkcí proteinu (shrnuto v Hoštáková, 2012). Významným krokem k pochopení a klasifikaci proteinových struktur je objasnění jejich sekundární struktury. Jedním z přístupů, jak toho docílit, je tzv. secondary structure *assignment (ASS)* - přiřazení konkrétního stavu sekundární struktury aminokyselině v 3D struktuře. V literatuře i v této práci se často používá také termín anotace sekundární struktury.

1.1 Proteiny a jejich struktura

Proteiny jsou polymery dvaceti různých aminokyselin, které se označují jako biogenní. (Petsko & Ringe, 2004) S 20 aminokyselinami (= AK) lze docílit ohromné variability. Pro dipeptid existuje 780 možných permutací (shrnuto v Whitford, 2005).

Na základě Zhangovy analýzy (J. Zhang, 2000) je průměrná délka proteinu u Archebakterií 270±9 AK, bakterie mají průměr 330±5 a eukaryota pak 449±25 aminokyselin. Pro tyto hodnoty je již množství možných kombinací vyšší než odhadovaný počet atomů ve vesmíru.



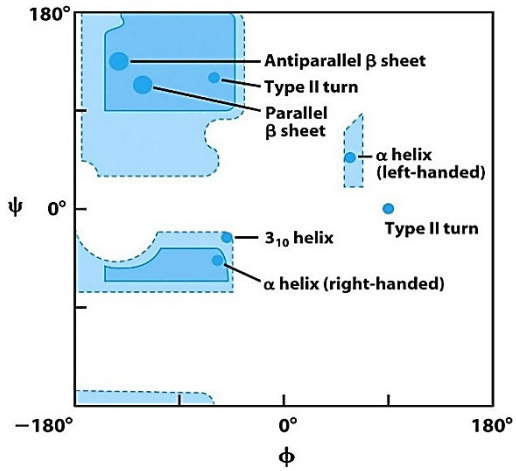
Obr. 1: Vlastnosti peptidové vazby. *R1, R2 a R3 představují postranní řetězce aminokyselin. Zbytek struktury je běžně označován jako proteinová kostra. Čtyřúhelníky značí rovinu, ve které se nacházejí atomy kostry. V rámci těchto rovin jsou konstantní vzdálenosti a úhly mezi atomy. Úhly Φ a Ψ jsou též znázorněny.*

Aminokyseliny jsou propojeny peptidovou vazbou (Petsko and Ringe 2004). Peptidová vazba je z chemického hlediska kovalentní vazbou, která je vytvořena mezi karboxylovou ky-

selinou a aminoskupinou za současné ztráty molekuly vody. Stabilita peptidové vazby je způsobena rezonancí. Tak se označuje delokalizace elektronů mezi několika atomy. Rezonance zvyšuje polaritu peptidové vazby, což může významně ovlivnit chování sbalených proteinů.

Dále má peptidová vazba díky rezonanci částečně charakter dvojné vazby, což znamená, že tři nevodíkové atomy tvořící vazbu (karbonylový kyslík, karbonylový uhlík a amidový dusík) jsou koplanární a volná rotace kolem vazby je omezená, viz **obr. 1**. Vazby N-Cα a Cα-C (k Cα je připojen postranní řetězec) jsou jednoduché a je možná volná rotace, není-li tam žádná sterická interference vyvolaná například postranními řetězci. Úhel vazby N-Cα k sousední peptidové vazbě je známý jako torzní úhel phi (Φ), zatímco úhel mezi vazbou C-Cα a sousední peptidovou vazbou je známý jako psi (Ψ), viz **obr. 1** (shrnuto v Petsko & Ringe, 2004).

Povolené hodnoty pro Φ a Ψ byly poprvé z energetického hlediska kalkulovány G. N. Ramachandranem (Ramachandran, Ramakrishnan, & Sasisekharan, 1963). Tyto hodnoty jsou zaneseny ve dvourozměrném grafu Φ / Ψ, který je nyní nazýván Ramachandranův **(Obr 2)**. Ramachandranův graf se výrazně liší pro glycin. Ten má díky absenci postranního řetězce nejvíce "povolených" hodnot (Φ / Ψ) (Petsko & Ringe, 2004).



Obr. 2: Ramachandranův graf *znázorňující typické dihedrální (Φ / Ψ) úhly různých SSE. Povolené oblasti jsou modře, sytě jsou znázorněny preferované hodnoty, světle naopak hraniční.*

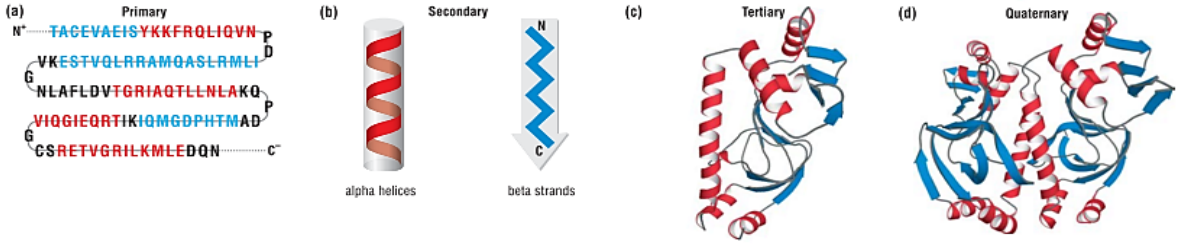
Struktura proteinů se běžně popisuje na čtyřech úrovních. Primární struktura odpovídá sekvenci aminokyselin. Jako sekundární struktura je označována lokální konformace polypeptidového řetězce nebo prostorový vztah aminokyselinových zbytků, které jsou v primární sekvenci blízko sebe (Whitford, 1961).

Terciární struktura pak odpovídá celkové 3D struktuře. Kvarterní struktura označuje asociaci více samostatných proteinových podjednotek do funkčního komplexu. Hierarchická klasifikace struktury proteinů je schematicky znázorněná na **Obr. 3** (Petsko & Ringe, 2004).

V této práci se budu zabývat sekundárními strukturami, proto se na jejich definici budu soustředit více než na jiné úrovně proteinové struktury.

Pauling a Corey pracující v Kalifornském technologickém institutu predikovali v roce 1951 existenci dvou periodických motivů v proteinových strukturách: α-helix a β-list (Pauling & Corey, 1951). Navrhli jejich ideální modely na bázi vodíkových vazeb uvnitř kostry. Ve významných ohledech jsou tyto modely správné a platí dodnes, včetně délek vazeb. Neuvažo-

vali však točivost helixů nebo možnost ohýbání β-listů (Eisenberg, 2003). Dnes víme, že existují i další typy pravidelných sekundárních struktur, ty však představují varianty jednoho z těchto základních motivů (Whitford, 1961).



Obr. 3: Hierarchie proteinové struktury. *(a) primární, (b) sekundární, (c) terciární, (d) kvartérní.*

Pauling a Corey představili své modely 7 let předtím, než byly struktury celých globulárních proteinů poprvé odhaleny rentgenovou krystalografií. První výsledky zaznamenal v roce 1958 John Kendrew u myoglobinu, kde je veškerá sekundární struktura helikální (Kendrew et al., 1958). Ke svému objevu tehdy řekl: *"nejpozoruhodnějším rysem molekuly je její složitost a nedostatek symetrie. Struktura je komplikovanější, než bylo předpovězeno jakoukoli teorií."*

Na rozdíl od pravidelné struktury nukleových kyselin, které mají pouze informační a replikační funkci, nepravidelnost proteinů je nutná pro plnění jejich rozmanitých funkcí – proteiny musí specificky rozpoznat mnoho tisíců různých molekul. Navzdory těmto požadavkům existují pravidelnosti ve vnitřní stavbě proteinových struktur, z nichž nejdůležitější je právě sekundární struktura. Sekundární struktury se ukázaly být hlavními rysy proteinové stavby. Tyto pravidelně se opakující makroelementy pozorujeme ve všech známých strukturách (Martin et al., 2005). Z analýz provedených o pět desetiletí později víme, že v průměru asi polovina zbytků v proteinech se účastní helixů či listů (Berman, 2000).

Důvodem, proč sekundární struktury vznikají, je stabilizace hydrofobního jádra proteinu. Kendrew si všiml, že aminokyseliny ve vnitřku proteinu mají téměř výhradně hydrofobní postranní řetězce. Vytvoření hydrofobního jádra z proteinového řetězce při procesu balení představuje určitou nesnáz. Aby se hydrofobní postranní řetězce dostaly do jádra, musí se dostat dovnitř také hlavní řetězec. Ten je však vysoce polární a proto hydrofilní. Obsahuje totiž jeden donor vodíkové vazby (NH) a jeden akceptor (C=0) na každý AK zbytek. V hydrofobním prostředí musí být tyto polární skupiny hlavního řetězce neutralizovány tvorbou vodíkových vazeb. To je řešeno vytvořením pravidelné sekundární struktury uvnitř molekuly proteinu (shrnuto v Brändén & Tooze, 1999).

Další úroveň hierarchie, která stojí mimo již zmíněné pojmy je *supersekundární struktura.* Jde o jakýsi přechod mezi sekundární a terciární strukturou. Jsou tak typicky označovány relativní vzdálenosti a vzájemná prostorová orientace mezi dvěma nebo více SSE. Stejné supersekundární struktury mohou někdy sdílet funkci. Do této kategorie patří například motiv "řeckého klíče", pojmenovaný podle podobnosti s řeckým výtvarným vzorem. Tento motiv se snadno tvoří během procesu skládání proteinů. Dalším příkladem supersekundárních struktur jsou DNA-vazebné motivy typu Zinc-finger asociované s ionty zinku (Andersen & Rost, 2005; Hutchinson & Thornton, 1993).

2 Typy sekundárních struktur proteinů

Existují tři hlavní kategorie sekundárních struktur: helix, list a coil. Mezi helixy patří nejběžnější α-helix, dále pak 3_{10}-helixy či π-helixy. Mezi beta struktury se typicky řadí β-hřeben (β-strand) - z několika β-hřebenů se skládá β-list (β-sheet). Coil motivy se dále člení na ohyby (turns) a smyčky (loops). Geometrické vlastnosti těchto struktur jsou shrnuty v **Tabulce 1**. Nejběžnějším způsobem třídění SSE je DSSP klasifikace (Kabsch & Sander, 1983). Ta vychází ze vzorů HB.

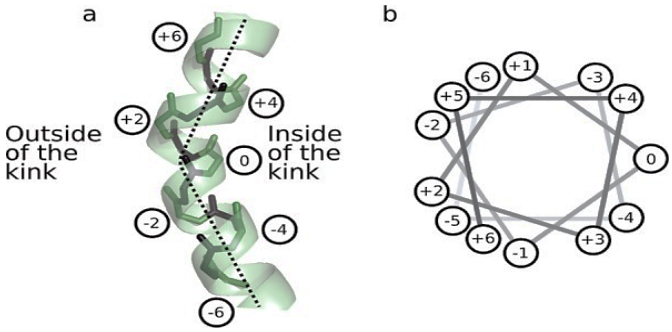
SSE	Φ	Ψ	Translation distance [nm]	Počet zbytků na otáčku
α helix	−57	−47	0.150	3.6
3_{10}-helix	−59	−26	0.200	3.0
π-helix	−57	−70	0.115	4.4
Poly(Pro) I	−83	+158	0.190	3.3
Poly(Pro) II	−78	+149	0.312	3.0
Paralelní β list	−139	+135	0.320	2.0
Antiparalelní β list	−119	+113	0.340	2.0

Tabulka 1: *Dihedrální úhly, translation distance na jeden zbytek a počet zbytků na otáčku pro pravidelné konformace sekundární konstrukce (Whitford, 1961).*

2.1 Helikální SSE

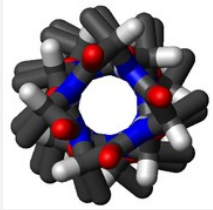
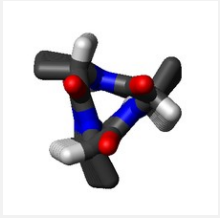
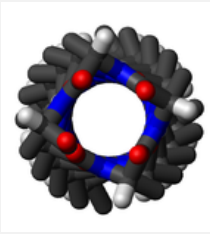
Mezi nejčastější struktury s helikální topologií se řadí α-helix, 3_{10}-helix a π-helix. Díky odlišným fyzikálně-chemickým parametrům existují mezi těmito typy helikálních struktur podstatné rozdíly z hlediska stability a interakčních schopností. Některé typické vlastnosti helixů jsou uvedeny v **Tabulce 2**.

Helixy jsou obvykle přímě, ale mohou se v nich vyskytovat i ohyby. Ohyby v helixech znázorněné na **Obr. 4 a)** jsou společným znakem proteinů s membránovými α-helixy, ale byly považovány za vzácné v solubilních proteinech. Law a kol. ukázali, že zalomené helixy nejsou specifické pro membránové proteiny Podobný počet ohnutí je viditelný v membránových i solubilních helixech podobné délky (Law et al, 2016).



Obr. 4: Ohyb v helixu: *a) Číslování pro příklad ohybu. Přerušovaná čára zobrazuje přibližnou osu helixu. b) Helix wheel diagram znázorňující helix v (a); Rezidua jsou očíslovány od N-konce k C-konci, zbytek v ohybu je označen číslem 0. Rezidua -4, -3, 0, +3 a +4 jsou tak na vnitřní straně ohybu, zatímco - 5, -2, +2 a +5 jsou na straně vnější.*

Dominantní AK jak v membránových, tak v solubilních zlomech helixů je prolin, ačkoliv více prolinů je zabudováno do dlouhých membránových helixů, než do dlouhých helixů solubilních proteinů (Law et al., 2016).

Geometrická vlastnost	**α-helix**	**3_{10} helix**	**π-helix**
Schéma			
Zbytků na otočku	3.6	3.0	4.4
Stoupavost na AK zbytek	1.5 Å	2.0 Å	1.1 Å
Poloměr helixu	2.3 Å	1.9 Å	2.8 Å
Stoupání	5.4 Å	6.0 Å	4.8 Å
Motiv HB	i → i+4	i → i+3	i → i+5
Úhly mezi sousedními AK	Ω = 100°	Ω = 120°	Ω = 87°

Tabulka 2: *Běžné typy helikálních struktur s jejich ideálními vlastnostmi. Stoupání = svislá vzdálenost mezi po sobě jdoucími otáčkami. Motiv HB = můstky vnikají mezi i-tým a i+n-tým reziduem (n = 3, 4, 5).*

Pro přehledné zobrazení helixů se běžně používá Helical wheel graf. Příklad takového zobrazení je na **Obr. 4 b)**. Graf zobrazuje trojrozměrnou strukturu α-helixu pomocí dvojrozměrné projekce. Tato reprezentace je vhodná například pro popis amfipatických helixů s jednou stranou s hydrofobními rezidui a druhou s hydrofilními. Toto uspořádání je běžné u α-helixů v globulárních proteinech, kde jedna plocha šroubovice je orientována k hydrofobnímu jádru a jedna plocha je orientována směrem k povrchu vystavenému rozpouštědlu (Schiffer & Edmundson, 1967). Může však sloužit k označení libovolné jiné vlastnosti reziduí.

Důležitou vlastností helixů je jejich točivost. Helixy rozdělujeme na pravo- a levotočivé v závislosti na směru "závitu" hlavního řetězce. L-aminokyseliny však většinou neumožňují levotočivé uspořádání helixu v důsledku těsného přiblížení postranních řetězců a skupiny CO. Naopak, D-aminokyseliny levotočivou konformaci preferují. Helixy, které jsou pozorovány v proteinech, jsou téměř vždy pravotočivé. Krátké levotočivé oblasti (3-6 zbytků) se vyskytují pouze příležitostně (Branden & Tooze, 1999).

Průměrné hodnoty dihedrálních úhlů levotočivých α- a 3_{10}-helixů jsou blízké typickým hodnotám pro pravotočivé, ale jsou s opačným znaménkem.

Navzdory jejich relativně nízkému zastoupení v proteinech se předpokládá, že levotočivé helixy hrají významnou roli ve funkci proteinů. Bývají například součástí vysoce konzervovaných aktivních míst enzymů, jsou zodpovědné za substrátovou specifitu i strukturní stabilitu nebo se podílí na vazbě iontů (Novotny & Kleywegt, 2005).

Jednotlivé typy helixů mohou za určitých podmínek přecházet v typy jiné. Bylo zjištěno, že α-helix prochází strukturálními přechody na π- nebo 3_{10}-helix, když je délka helixu napnuta o více než 10% oproti relaxovanému stavu. Bariéry pro strukturální přechody spojené hlavně s

přerušením vodíkových vazeb jsou v polyalaninu značně ovlivněny bočními řetězci. Tento efekt nemůže být přičítán pouze odpuzujícím interakcím mezi bočními řetězci a hlavním řetězcem helixu, ale významným změnám v kovalentních vazbách v peptidové jednotce polyalaninu oproti polyglycinu (Ireta, 2018).

2.1.1 α-helix

α helix (3.6_{13}-helix) je nejběžnější strukturní motiv nacházející se v proteinech; v globulárních proteinech tvoří okolo 31% (Fodje & Al-Karadaghi, 2002). Tento SSE se nejsnáze anotuje. Název α získala tato struktura díky tomu, že Paulingův návrh byl zčásti založen na datech rentgenové difrakce vláknitého proteinu α-keratinu (Formaggio et al., 2013).

Stojí za zmínku, že α-helix není charakterizován celočíselným počtem aminokyselin na otočku (3.6). Pauling musel proto vyvrátit obecný názor strukturních biochemiků na počátku padesátých let, aby jeho návrh přijala vědecká komunita. V té době byly totiž považovány za stabilní pouze helixy s celočíselným počtem aminokyselin na otáčku.

Ideální α-helix má pravidelně vazby mezi rezidui i → i + 4 a končí dvěma po sobě následujícími i - 4 ← i vodíkovými vazbami. Podle DSSP klasifikace jde o 4-turn, stav je značený písmenem H. Díky tomuto vzoru jsou všechny skupiny NH a CO spojeny s vodíkovými vazbami, vyjma prvních NH skupin a posledních CO skupin na koncích helixu. V důsledku toho jsou konce helixů polární a jsou téměř vždy na povrchu proteinových molekul. Kvůli polaritě peptidové vazby je amino konec α-helixu pozitivní, zatímco karboxylový konec je negativní. Tyto náboje by měly přitahovat ligandy opačných nábojů. Negativně nabité ligandy, zejména pokud obsahují fosfátové skupiny, se často váží na N-koncích helixů. Naproti tomu pozitivně nabité ligandy se na C-konci váží jen zřídka. Jde o příklad specifické vazby na bázi konformace hlavního řetězce, ve které nehrají roli postranní řetězce (shrnuto v Brändén & Tooze, 1999; Whitford, 2005).

Vzhledem k tomu, že v α-helixu je ~ 3,5 aminokyselin na otáčku, nejmenší celé číslo charakterizující tento helix je 7, což odpovídá dvěma úplným otočkám. To je důvod, proč jsou biologicky relevantní amfipatické helixy charakterizovány opakováním sedmic aminokyselin *(a, b, c, d, e, f, g)* s analogickými fyzikálně-chemickými vlastnostmi na specifických pozicích v rámci této sedmice. Například ve vodných roztocích vyžadují polohy *a, d* hydrofobní zbytky, zatímco pro polohy *e* a *g* jsou vyžadovány zbytky hydrofilní (Formaggio et al., 2013).

Vodíkové vazby od atomu kyslíku k atomu dusíku jsou u ideálního modelu dlouhé 0,286 nm, jsou lineární, míří jedním směrem a leží paralelně k ose helixu. Peptidové jednotky jsou podél osy stejně orientované. U reálných struktur existují mírné odchylky v délce můstků i hodnotách vazebných úhlů (Branden & Tooze, 1999; Whitford, 1961).

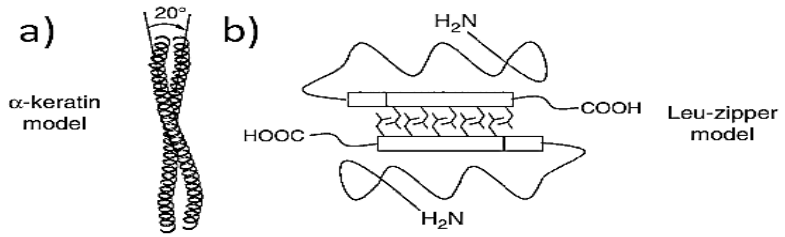
První čtyři skupiny NH a poslední čtyři skupiny CO mají normálně nedostatečné vodíkové vazby. Z tohoto důvodu mají velmi krátké helixy často zkreslené konformace a tvoří alternativní partnery vodíkových vazeb.

Deformace vodíkových vazeb a délky, k nimž dochází v reálných helixech, jsou doprovázeny hodnotami dihedrálních úhlů (Φ a Ψ), které se významně odlišují od ideálních hodnot -57◦ a -47◦ (Whitford, 1961).

Délka α-helixů v globulárních proteinech se výrazně liší. Nejkratší jsou jen 4-5 aminokyselin dlouhé, existuje ale i helix dlouhý 40 aminokyselin. Průměrná délka je přibližně deset zbytků,

což odpovídá třem otáčkám. Stoupavost helixu podél osy na jeden zbytek (transition per residue) je 1,5 Å, což odpovídá průměrné délce helixu přibližně 15 Å (Branden & Tooze, 1999; Whitford, 1961).

Dva α-helixy mohu vzájemně asociovat a vytvořit tak takzvanou “coiled-coil” strukturu, viz **Obr. 5 a)**, kdy jsou dva helixy vzájemně propletené, s úhlem mezi osami přibližně 20°. Poprvé byla tato struktura objevena na α-keratinu v roce 1952 (Crick, 1952). Pro Leu-bohaté helixy byl tento motiv v roce 1988 McNightem nazván "Leu zip", viz **Obr. 5 b)** a je jedním z rozšířených DNA-vazebných proteinových motivů (Formaggio et al., 2013; Landschulz, Johnson, & McKnight, 1988).



Obr. 5: *α-helikální motivy. a): coiled-coil motiv, b): Leu-zip motiv* (Alemán, Bianco, & Venanzi, 2013).

2.1.2 3_{10}-helix

Tento typ SSE přilákal pozornost strukturních biochemiků a proteinových krystalografů relativně později, ačkoliv teoreticky byla tato struktura popsána dříve než α-helix, a to již v roce 1941 (Taylor, 1942)**.** Důvodem je, že 3_{10}-helix je charakterizován celočíselným počtem (3) aminokyselin na otočku (Formaggio et al., 2013).

Často se vyskytuje v proteinech, pokud je normální α-helix narušen přítomností nepříznivých AK zbytků, dále v blízkosti ohybů nebo když se krátké sekvence skládají do helikální konformace (Whitford, 1961). Dále se ukázalo, že Cα-methylované AK mají tendenci podporovat vznik 3_{10}-helikální struktury (Formaggio et al., 2013; Toniolo et al., 2001). 3_{10}-helixy mohou stejně jako π-helixy tvořit “rozšiřující” C-terminální nebo N-terminální konce α-helixů o délce 3-4 rezidua. Přestože naprostá většina 3_{10}-helixů je krátká, byly nalezeny i úseky o délce 7–12 reziduí.

Pavone a kol. byli schopni jako první připravit a analyzovat pravidelný 3_{10}-helix o délce tří úplných otoček. Krystalizovali syntetický homodekapeptid z alfa-alfa-dimethylovaného glycylového zbytku alfa-aminoizomáselné kyseliny do podoby jediného krystalu. Tento krystal byl analyzován metodou X-ray difrakce. Sloučenina krystalizuje jako perfektní 3_{10}-helix, stabilizovaná osmi po sobě následujícími intramolekulárními N-H...O=C vodíkovými můstky (Pavone et al., 1990).

Označení 3_{10} odkazuje na počet atomů kostry umístěných mezi atomy donoru a akceptoru vodíkové vazby (10) a skutečnost, že existují tři zbytky na otočku (Whitford, 1961). Stav v DSSP klasifikaci se značí G a jde o 3-turn, tedy strukturu s i → i + 3 motivem (Kabsch & Sander, 1983). Jde o celkově pevnější a užší strukturu, než jakou je α-helix. Je zde vyšší potenciál pro nepříznivé kontakty mezi atomy kostry nebo postranních řetězců (Whitford, 1961). 3_{10}-helix není, podobně jako π-helix, energeticky výhodná konformace, jelikož atomy kostry jsou u 3_{10}-helixu příliš těsně u sebe.

Pauling a kol. nesprávně předpovídali, že se v proteinech tato struktura nevyskytuje právě v důsledku nepříznivých vazebných úhlů. V této konformaci jsou však pozorována přibližně 4% všech aminokyselin (Fodje & Al-Karadaghi, 2002), což odpovídá 10% všech helikálních residuí v globulárních proteinech (Barlow & Thornton, 1988; Formaggio et al., 2013).

Pro stav koncových AK hraje roli přítomnost okolních SSE. Aminokyselina v koncových polohách α-helixu je součástí jiného SSE v závislosti na konkrétním kontextu. Pokud existuje sousední 3_{10}-helix, vzniká složená šroubovice.

Pal a kol. (Pal, Dasgupta, & Chakrabarti, 2005) provedli analýzu 138 případů 3_{10}-alfa a 107 alfa-3_{10} složených helixů, nacházejících se ve známých proteinových strukturách. Výsledky naznačují, že SSE, který se vyskytuje první, ukládá své charakteristiky následujícímu SSE. Když předchází 3_{10}-helix, tendence prolinu nacházet se v poloze N1 alfa-helixu je posunuta do polohy N2, což je typická charakteristika C-koncového 3_{10}-helixu. U druhého typu složeného helixu vzniká mezi SSE ohyb, přičemž dva spojovací AK zbytky směřují dovnitř a jsou zanořeny do struktury. Složené helixy tohoto typu tedy mohou být lomené. Podobný typ terciární struktury zaujímá i α-helix v místě vloženého prolinu.

2.1.3 π-helix

Jedná se o další z helikálních struktur, jejíž výskyt není v proteinech tak častý. Krátké π-helixy se nacházejí v 15% známých proteinových struktur (Cooley, Arp, & Karplus, 2010).

Pokud se budeme držet DSSP klasifikace, jedná se o pravidelný motiv s vazbami i → i + 5, což je označováno jako 5-turn (stav I). K formování takového vzoru HB je třeba zarovnat 5 AK, což je entropicky nákladnější. To je jeden z důvodů, proč je tento typ helixu považován za méně stabilní.

Tato struktura byla teoreticky popsána a pojmenována R. B.Baybuttem roku 1952, 4.4_{16}-helix je její alternativní označení (Baybutt, 1952). Existuje hypotéza, že jde o evoluční adaptaci odvozenou vložením jedné aminokyseliny do α-helixu (Cooley et al., 2010)

π-helix má 4.4 AK zbytky na otočku. Každý čtvrtý zbytek je v téměř azimutální poloze vzhledem k prvnímu. Postranní řetězce v π-helixu jsou od sebe nejblíž ve srovnání s jinými typy helixu (Fodje & Al-Karadaghi, 2002).

Další důvod nestability představuje fakt, že atomy jsou v π-helixu zabaleny tak volně, že jejím středem v ose helixu prochází dutina. Ještě, než byl experimentálně ověřen výskyt π-helixu, byla kvůli tomu tato konformace odmítána. Tato dutina je užší, než u gama helixu (viz dále), a není dostatečně velká pro umístění molekuly vody. Neexistuje tedy způsob, jak překlenout delší vzdálenosti napříč dutinou pomocí slabých interakcí (Baybutt, 1952). Vzhledem ke krátké vzdálenosti mezi postranními řetězci však může docházet ke stabilizaci interakcemi mezi nimi, jak poukázali (Fodje & Al-Karadaghi, 2002). Tyto interakce jsou většinou typu van der Waals, ale také jde o stahovací (stacking) interakce mezi aromatickými kruhy.

Dihedrální úhly předpokládaného dokonalého π-helixu jsou se svými hodnotami Φ a Ψ rovnými –57°a –70° na samé hranici povolených hodnot v Ramachandranově grafu. MD simulace nicméně ukazují spíše na hodnoty kolem –77° a –54°.

π-helix může, navzdory své předpokládané nestabilitě, mít některé funkční výhody oproti jiným helikálním strukturám. Postranní řetězce funkčně důležitých zbytků v každé čtvrté pozici uvnitř π-helixu mohou být poskládány blízko sebe a spojeny způsobem, který není umožněn

v žádném jiném typu helixu. Navíc zřejmě existují mechanismy, díky kterým je π-helix stabilnější, než by bylo na první pohled patrné (Fodje & Al-Karadaghi, 2002). Autoři ukázali, že π-helix má specifické preference výskytu AK. Navíc je π-helix konzervován ve funkčně příbuzných proteinech. Tyto preference jsou dále diskutovány v popisu metody SECSTR.

Ne všechny teoreticky předpovězené helikální konformace byly objeveny v reálných strukturách. Příkladem může být již zmíněný gama helix, který byl také předpovídán Paulingem a Coreym jako možný konstrukční prvek v proteinech. Tato šroubovice, mající označení 3.6_{14}, nebyla však nikdy pozorována (Berndt, 1996).

Další třídou jsou polyprolinové (Poly(Pro) I, Poly(Pro) II) helixy. Druhý typ je levotočivý a nachází se často v přirozeně nesbalených proteinech. Také hraje klíčovou roli v interakcích mezi proteiny a nukleovými kyselinami a mezi proteiny navzájem (Adzhubei, Sternberg, & Makarov, 2013).

Ostatní typy helikálních struktur, jako je 2_{27}-helix, nejsou v proteinech tak časté.

2.2 Beta struktury

β-struktury představují druhý základní typ běžně se vyskytujících SSE. Nejběžnější jsou β-list, poskládaný ze základních jednotek - β-hřebenů. Dalším ze specifických typů beta struktur je β-Barel. Jde o uzavřený kruh paralelních β-hřebenů vzájemně pospojovaných helikálními segmenty (Whitford, 1961).

2.2.1 β-Hřeben

Navzdory svému názvu je β-hřeben též helixem, i když extrémně protáhlou formou s dvěma AK zbytky na otočku. Jeden β-hřeben není stabilní, převážně kvůli omezenému počtu lokálních interakcí. K výrazné stabilizaci dochází propojením dvou a více těchto motivů do struktury β-Listu.

'Cartoon' reprezentace β-hřebenů zjednodušuje rozpoznání směru hlavního řetězce v molekulárních strukturách. β-hřebeny jsou běžně znázorněny jako šipky, vedoucí směrem od N k C konci.

Na rozdíl od ideálních modelů jsou reálné β-hřebeny často zkresleny zkroucením, kvůli pozitivnějším hodnotám (φ / ψ) úhlů. Výsledkem je nepatrná, ale rozpoznatelná pravotočivost polypeptidového řetězce (Whitford, 1961).

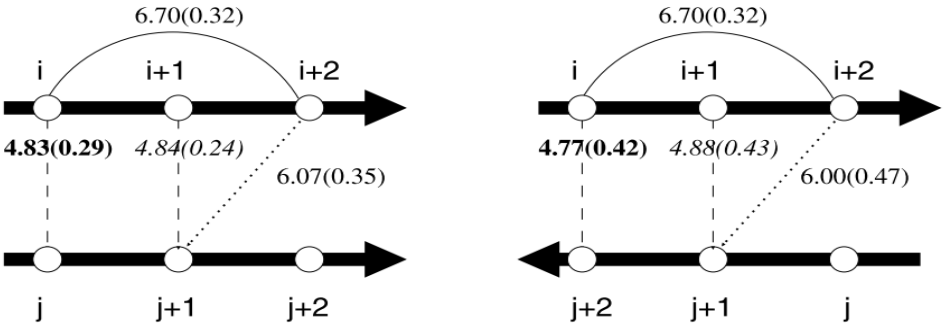
2.2.2 β-List

Helix je vytvořen z jedné spojité oblasti polypeptidovédho řetězce. Naproti tomu struktura β-listu (známá též pod pojmem beta skládaný list) je sestavena z kombinace hned několika nesousedících oblastí - β-hřebenů. Tyto β-hřebeny mají obvykle délku 5 až 10 AK (Branden & Tooze, 1999). Vzhledem k této komplexnější stavbě je relativně těžší tyto struktury anotovat.

β-hřebeny mohou vzájemně interagovat dvěma způsoby, aby vytvořily skládaný list. Aminokyseliny mohou v přilehlých β-hřebenech ležet všechny ve stejném biochemickém směru (amino konec a karboxylový konec). V tomto případě je list popsán jako paralelní. Nebo mohou mít aminokyseliny v přilehlých řetězcích střídavé směry, pak se list nazývá antiparalelní. β-hřebeny mohou být také spojeny do smíšených β-listů, ovšem pouze okolo 20% β-hřebenů

uvnitř β-listů známých proteinových struktur má na jedné straně paralelní vazby a antiparalelní vazby na straně druhé (shrnuto v Brändén & Tooze, 1999).

Typické vzdálenosti v β-listu jsou znázorněny na **Obr. 6**.



***Obr. 6: Typická vzdálenost Cα v β-listech**. Typické vzdálenosti Cα paralelních (levá část) a antiparalelních β-listů. Průměrné vzdálenosti jsou uvedeny v Å s jejich standardními odchylkami v závorkách. Samostatná statistika byla vypočtena pro vzdálenosti zahrnující pouze zbytky v jádrech vlákna (kurzívou) a vzdálenosti, které obsahují zbytky na koncích (tučně). Vzdálenost v rámci hřebenu (i, i+2) nezávisí na jeho orientaci v listu. (Martin et al., 2005).*

Antiparalelní listy jsou častější, protože jsou často tvořeny pouze ze dvou opačně orientovaných β-hřebenů. U paralelních listů je pozorována interakce alespoň čtyř β-hřebenů (Whitford, 1961).

2.2.3 β-Turn

Nejjednodušší sekundární strukturní prvek, který obvykle zahrnuje čtyři AK, ale někdy vyžaduje pouze tři. Vodíkový můstek vzniká mezi i a i+3 reziduem. Tento vzorec vodíkových vazeb nemůže dále pokračovat, protože je otočka příliš těsná. Turn způsobuje změnu směru polypeptidového řetězce. Většina skupin C=O a N-H ve čtyřech AK zbytcích tvořících turn nevytváří HB s jinými atomy kostry.

β-turns se nacházejí na povrchu složených bílkovin, kde jsou v kontaktu s vodním prostředím, a obrácením směru řetězce mohou omezit velikost molekuly a udržovat její kompaktní stav (Petsko & Ringe, 2004).

V některých proteinech může být podíl AK nacházejících se v obrátkách vyšší než 30%.

AK s objemnými či rozvětvenými postranními řetězci se v obrátkách vyskytují ve velmi nízkých frekvencích. Častěji jsou zde zastoupeny AK zbytky s malými postranními řetězci, jako je Gly, Asp, Ser, Cys a Pro (Whitford, 1961).

2.3 Smyčky (Loops)

Oblasti smyček jsou převážně na povrchu molekuly a jsou bohaté na nabité a polární hydrofilní AK zbytky. CO a NH skupiny hlavního řetězce těchto oblastí netvoří navzájem HB. Jsou tak vystaveny solvataci a mohou vytvářet HB s molekulami vody. Smyčky jsou oproti jádrům globulárních proteinů mnohem méně evolučně konzervované. Většina mutací probíhá v těchto oblastech. Častou funkcí smyček je spojování ostatních SSE jednotek. Mimo to se

tyto oblasti účastní tvorby vazebných míst a enzymových smyčkových oblastí. Variabilita smyček co do délky i sekvence je veliká.

Smyčkové oblasti, které spojují dvě sousední antiparalelní 3D struktury, mají spíše omezené spektrum konformací. Nazývají se harpin smyčky. Krátkým harpin smyčkám říkáme turns. Nejčastější jsou dva typy: Type I turn a Type II turn. Druhý typ obvykle obsahuje Gly jako druhý ze dvou vnitřních AK zbytků.

Dlouhé smyčkové oblasti jsou často flexibilní a mohou zaujímat několik konformací, což je činí "neviditelné" při stanovení struktury. Takové smyčky hrají často roli ve fungování proteinu a mohou se měnit z "otevřené" konformace, která umožňuje přístup k aktivnímu místu, do "uzavřené" konformace, která chrání reaktivní skupiny v aktivním místě před vodou.

Jeden konkrétní typ dlouhé smyčky, zvaný omega smyčka podle svého tvaru, je kompaktní a má stabilní vnitřní interakce. Jde tak o jednu z nejstabilnějších smyček. Ostatní dlouhé smyčky jsou často stabilizovány interakcemi s ionty kalcia (Branden & Tooze, 1999).

3 Význam sekundární struktury

Sekundární struktury umožňují jednoduchý a intuitivní popis 3D struktur. Sekundární struktury jsou široce využívány pro strukturní i sekvenční porovnání a strukturní klasifikaci. V neposlední řadě poskytují přirozený způsob pro vizualizaci struktury (Martin et al., 2005).

Příkladem využití sekundární struktury pro zlepšení multiple sequence alignmentu je PRALINE toolbox (Bawono & Heringa, 2014; Simossis & Heringa, 2005). Ten kromě řady strategií alignmentu nabízí výběr ze sedmi různých predikčních metod sekundární struktury, které jsou použitelné samostatně či v kombinaci za účelem integrace strukturních informací do procesu alignmentu. PRALINE kromě toho nabízí i možnost integrovat informace získané z vyhledávání homologů.

Charakteristické rysy dané prostorové konfigurace (foldu) bývají při klasifikaci odborníky často popsány právě celkovým vzájemným uspořádáním sekundární struktury. Její znalost je tedy pro strukturní klasifikaci nutná (Andersen & Rost, 2005; North, 1992).

Správné určení sekundární struktury je také důležitým krokem pro metody komparativního modelingu (Chothia & Lesk, 1986) a threadingu (Bowie, Luthy, & Eisenberg, 1991). Jde o dva odlišné přístupy predikce 3D struktury. Komparativní modeling využívá fragmenty existujících 3D struktur jako templáty pro určení struktury predikovaného proteinu. Základní myšlenka threadingu je co nejlépe “navléct” určitou strukturu na danou sekvenci. Obě tyto metody jsou důležité pro určení oblastí proteinu, které mohou například tvořit aktivní místa enzymů.

Sekundární struktura se používá v hierarchické klasifikaci pro určení strukturní třídy, která kategorizuje proteiny v závislosti na množství a uspořádání sekundárních struktur. SCOP (= structural classification of proteins) je představitelem těchto metod klasifikace (Hubbard et al., 1999). Ten třídí proteiny na čtyřech úrovních vzájemné podobnosti. Dalším z nich je CATH (= Class, Architecture, Topology, Homologous superfamily) (Orengo et al., 1997), který je na rozdíl od SCOP plně automatický a je proto se vzrůstajícím počtem známých struktur používaný stále častěji.

3.1 Databáze a SSE

Výše zmíněné metody SCOP a CATH tvoří metody kategorizace proteinů. Současně se jedná o plnohodnotné databáze, které tyto metody používají. SCOP databáze poskytuje podrobný a komplexní popis strukturních a evolučních vztahů proteinů známé struktury.

Kromě toho existuje i specializovaná databáze sekundárních strukturních elementů: PSS = Protein Secondary Structure (Suzuki et al., 1991). Jsou zde zobrazeny oblasti s definovanými sekundárními strukturami a oblasti strukturního zájmu společně s primárními strukturami proteinů. Na základně dotazu je možné zobrazit sekundární strukturu požadované délky peptidového fragmentu, stejně jako peptidový fragment, který odpovídá definované sekundární struktuře. Tato databáze navíc obsahuje software, který identifikuje páry AK, které mají mezi sebou HB a počítá frekvenci výskytu každé dvojice, stejně jako parametry konformace.

Kromě samotných databází potřebujeme i efektivní nástroj, jak v databázi vyhledávat. K tomu obecně slouží dotazovací jazyky, z nichž nejvíce používaným je SQL. Ten je oblíbeným nástrojem pro správu relačních databází. Bohužel ale není navržen pro účinné uchovávání a zpracování biologických dat, jako jsou sekundární struktury bílkovin.

Existuje nicméně PSS-SQL (Mrozek et.al., 2010), což je rozšíření SQL pro biologická data. Umožňuje formulovat dotazy na databázi tak, aby nalezly proteiny mající sekundární struktury podobné strukturním vzorům specifikovaným uživatelem. Je tak vhodným nástrojem pro vyhledávání podobnosti proteinové struktury. PSS-SQL je integrován do systému správy relačních databází *Microsoft SQL Server* a zjednodušuje manipulaci s biologickými daty.

4 Cíle práce

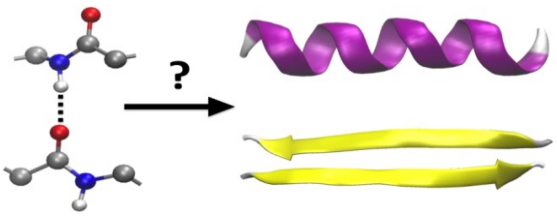
Cílem práce je udělat literární rešerši dostupných metod pro anotaci sekundární struktury na základě 3D strukturních dat a popisu SSE jako takových. Speciální pozornost bude věnována vzácnějším typům sekundárních struktur, jako jsou 3_{10}-helixy či π-helixy. Tato práce bude sloužit jako teoretický základ pro budoucí implementaci algoritmu pro anotaci do nástroje MolQL.

5 Metody anotace SSE

Výše byly popsány charakteristiky jednotlivých SSE na základě jejich různých vlastností. Nyní, když máme jednotlivé SSE dobře definovány, přistupme obecně k problematice jejich anotace a dále k jednotlivým metodám anotace.

5.1 Problematika anotace

K sekundární struktuře se dá dostat dvěma naprosto odlišným způsoby. Pokud máme pouze informaci o sekvenci a na jejím základě se snažíme "uhodnout" sekundární strukturu, pak hovoříme o predikci. Pokud vycházíme ze strukturních dat, hovoříme o *assignmentu* (anotaci). V této práci se budu zabývat téměř výhradně anotací, byť predikce je také velmi zajímavé téma. Úkolem anotace je přiřadit každé AK stav, popisující její sekundární strukturu na základě známé 3D struktury, viz **Obr. 7**.



Obr. 7: Schéma anotace.
Nalevo primární, napravo sekundární struktura proteinu (Haghighi, Higham, & Henchman, 2016).

Zdrojem dat pro anotaci je soubor 3D struktury proteinu, nejčastěji získaný z databáze PDB (= Protein Data Bank). Nepovinnou součástí PDB souborů je pro nás důležitá informace o sekundární struktuře. Ta je uložena v polích HELIX, SHEET a TURN. Přibližně 90% souborů PDB tato pole obsahuje. Nicméně, i když jsou tato pole používána, může se stát, že jen pár sekundárních strukturních elementů, které jsou pro vkladatele zajímavé, je popsáno a ostatní jsou ignorovány (Martin et al., 2005).

ASS není jednoduchý úkol. Jedním z důvodů je rozdíl mezi reálnými a ideálními sekundárními strukturami. Reálné sekundární struktury jsou náchylné k ohýbání, zkroucení, a dalším deformacím, čímž se liší od geometricky pravidelných prototypů. V globulárních proteinech tedy postrádají sekundární struktury pravidelnost. Většina α-helixů je zde například jemně zakřivená. (Law et al., 2016). Je nutné počítat s možností určité míry odlišnosti, jak z hlediska vzorů vodíkových vazeb, tak i hodnot dihedrálních úhlů. Rozlišení vstupního 3D modelu má přímý vliv na kvalitu výsledného ASS. Kromě toho jsou proteiny dynamické a u NMR struktur (kde je zdrojem dat nukleární magnetická rezonance) máme často několik modelů, které reflektují variabilitu, a to i na úrovni sekundární struktury.

Zpočátku krystalografové přiřazovali sekundární strukturu vizuálně na základě 3D struktur. Byla to tenkrát jediná možnost, jak ASS provést a specialisté se občas neshodli. To vedlo k vývoji automatických metod (Andersen & Rost, 2005). První implementaci takových metod představili Levit a Greer v roce 1980. Algoritmus byl založen hlavně na torzních úhlech mezi Cα atomy (Levitt, 1980). Pak v roce 1983 představili svůj algoritmus Kabsch and Sander. Ten

je dnes označován zkratkou DSSP a zůstává jakýmsi “zlatým standardem” v oboru (Kabsch & Sander, 1983). Následně byly vyvinuty další metody.

Přehled nejčastěji používaných metod je uveden v **Tabulce 3**. Jejich fungování a specifikace jsou diskutovány níže. V dalším textu budu odkazovat na jednotlivé metody pomocí jejich běžně používaných názvů, citace jsou uvedeny v tabulce.

Název metody	**Citace**	**# cit. (WoS)**	**Princip**	**Helix**			**ß struktura**		**Coil**	
				α	3_{10}	π	List	Bulge	Loop	Turn
DSSP	(Kabsch & Sander, 1983)	9645	HB vzory	✓	✓	✓	✓	✓	✓	✓
DEFINE	(Richards & Kundrot, 1988)	307	Cα vzdálenosti	✓	✓	✗	✓	✓	✓	✓
STRIDE	(Frishman & Argos, 1995)	1485	(Φ/Ψ) úhly	✓	✓	✓	✓	✓	✗	✓
KAKSI	(Martin et al., 2005)	76	(Φ/Ψ) úhly + Cα vzdálenosti	✓	✓	✓	✓	✗	✗	✗
P-SEA	(Labesse, Colloc’h, Pothier, & Mornon, 1997)	81	Cα stopa		✓			✓		✓
STALE	(Haghighi et al., 2016)	0	HB vzory	✓	✓	✓	✓	✓	✓	✓
SECSTR	(Fodje & Al-Karadaghi, 2002)	138	HB vzory	✓	✓	✓	✓	✓	✗	✓
P-CURVE	(Sklenar, Etchebest, & Lavery, 1989)	106	Geometrie kostry	✓	✓	✓	✓	✗	✗	✓
ScrewFit	(Kneller & Calligari, 2006)	17	Cα atomy, lokální helikální parametry	✓	✓	✓	✓	✗	✓	✓

Tabulka 3: Srovnání schopností různých metod identifikovat jednotlivé typy SSE. *Sloučené buňky znamenají, že metoda přiřazuje dané SSE jako jeden stav. (Například nerozlišuje mezi jednotlivými typy helixů.) Podtržené jsou SSE, na které se metoda specializuje. Počty citací jsou brány z databáze Web of Science a jsou aktuální k datu 2.5. 2018.*

Protože ASS metody často využívají HB pro anotaci, v kapitole 5.2 následuje krátký přehled používaných definic HB.

5.2 Definice SSE pomocí vodíkových můstků

Vodíková vazba byla popsána Paulingem roku 1928 (Pauling, 1928). Patří společně s Van der Waalsovými silami do kategorie takzvaných nevazebných interakcí. Jde o nejsilnější z nich.

Bohatá síť vodíkových vazeb se vytváří ve vodě. Vzniká tak prostředí, ve kterém se polární molekuly účastní slabých interakcí, zatímco nepolární molekuly tuto síť narušují. Narušení mají za následek chybějící vazby, což způsobuje relativní vzrůst energie, ve srovnání se stavem, kde vazba nechybí. Zvýšení energie lze zabránit nebo ho alespoň minimalizovat balením nepolárních molekul k sobě.

Tento hydrofobní efekt je nejdůležitější silou při spontánním formování sekundární i terciární struktury. Balení nepolárních AK zbytků v hydrofobním jádře proteinu zahrnuje proces zanoření atomů polární kostry dovnitř a přerušení vodíkových vazeb s vodou. Aby se překonal tento energeticky nevýhodný stav, jsou polarity v rámci kostry spárovány. V jádře bílkovin se tak vytvoří vodíkové vazby, čímž se fixuje jejich konformace.

Hodnoty energie vodíkové vazby se v proteinech pohybují okolo -2 kcal·mol^{-1}. Jsou tedy řádově 10× slabší než iontové nebo kovalentní vazby. Stabilita sekundárních struktur je dána tím, že zde HB vznikají ve velkém množství. Asi 90% všech C=O a NH skupin kostry se účastní vodíkové vazby (Baker & Hubbard, 1984). Pro popis vodíkové vazby se používají minimálně tři modely.

Baker a Hubbard definovali roku 1984 vodíkové vazby podle velikosti úhlu (NHO = θ) mezi akceptorem a donorem vodíkové vazby a jejich vzdálenosti (rHO). Vodíková vazba je přiřazena, pokud platí: θ > 120◦ a rHO < 2,5 Å. Tento relativně hrubý model přežil poměrně dlouho, jelikož v době určování sekundární struktury vizuálně nebyla přesnější definice nutná.

Coulombova vodíková vazba

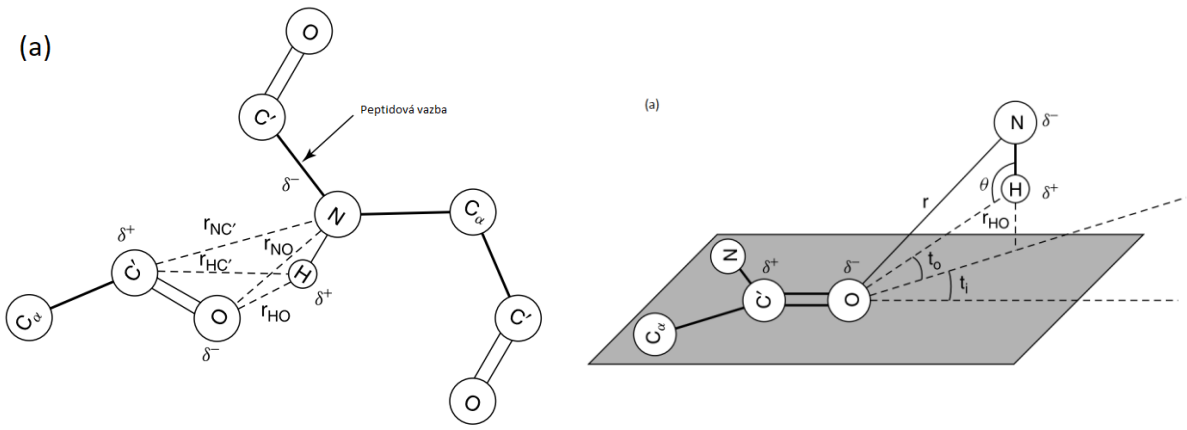
Novější model vodíkové vazby bere v potaz Coulombovu energii počítanou ze vzdáleností atomů. Pro každý pár aminokyselin se dá Coulombova energie vazby spočítat podle **Rovnice (1)** jako:

$$E = q_1 q_2 \left(\frac{1}{r_{ON}} + \frac{1}{r_{CH}} - \frac{1}{r_{OH}} - \frac{1}{r_{CN}}\right) \cdot 332\ kcal/mol \qquad (1),$$

kde q_1 a q_2 (označované též δ^+ a δ^-) jsou polární náboje uvedené v jednotkách elementárních elektronových nábojů *e*. r_{AB} pak označuje vzdálenost mezi atomy A a B. Vzájemné prostorové uspořádání atomů a jejich vzdálenosti jsou znázorněny na **obr. 8 a)**.

Nutná informace o pozici vodíkových atomů není v PDB souborech běžně uvedena, což vyžaduje její extrapolaci.

Coulombův energetický model nezahrnuje síly odpuzování atomů a neříká nic o charakteristické délce vodíkové vazby.



Obr. 8: *Výpočet energie vodíkové vazby. (a) Vzdálenosti používané pro výpočet Coulombovy vodíkové vazby, (b) Úhly a vzdálenosti definující empirickou vodíkovou vazbu.*

Empirická vodíková vazba

Poslední zde zmíněnou možností je výpočet energie empirické vodíkové vazby. Ten může být odvozen z geometrie HB v krystalických strukturách nebo z polypeptidů, peptidů, AK a malých organických sloučenin (Wade & Goodford, 1993)**.**

Celková energie E_{hb} závisí na energii vzdálenosti (NO) označované Er a na třech vazebných úhlech θ, t_i a t_0. Schéma úhlů a vzdáleností používaných pro výpočet empirické vodíkové vazby jsou schematicky znázorněny na **obr. 8 b)**. Rovnice pro výpočet E_{hb} má tvar:

$$E_{hb} = E_r \cdot E_t \cdot E_{p''}$$

Vztah pro Er je podobný Lennard-Jonesově potenciálu pro van der Waalsovu interakci, ale používá odlišné mocniny. Ep = cos2(θ) a Et má předpis:

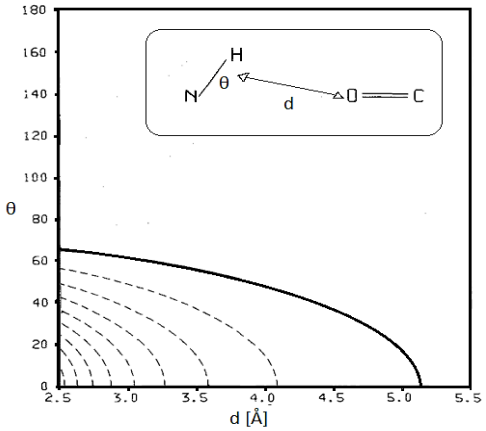
$$Et = \begin{cases} [0.9+0.1\sin(2t_i)]\cos(t_o) & 0^\circ < t_i \leq 90^\circ \\ K1[K2 - \cos2(t_i)]\cos(t_o) & 90^\circ < t_i \leq 110^\circ \\ 0 & 110^\circ \leq t_i \end{cases}$$

5.3 DSSP

Algoritmus DSSP (Dictionary of Secondary Structure of Proteins) je jedním z nejvíce používaných algoritmů pro anotaci SSE a je také jedním z nejstarších (Y. Zhang & Sagui, 2015). Dokladem jeho širokého využití může být fakt, že původní článek má v průměru 400 citací ročně (Haghighi et al., 2016).

DSSP patří do kategorie algoritmů určujících sekundární struktury na základě vzorů vodíkových můstků v rámci proteinové kostry. K výpočtu vodíkové vazby se zde používá Coulombův model HB. Atomy vodíku, které obvykle u struktur určených rentgenovou krystalografií chybí, jsou algoritmem přidány (Haghighi et al., 2016).

Parametry pro ideální vodíkový můstek podle DSSP jsou: d = 2.9 Å, θ = 0°, E = -3 kcal·mol^{-1}, přičemž DSSP přiřazuje vodíkové můstky až do hraničních hodnot E <= *-0.5 kcal·mol^{-1}*, θ <= 63° a d <= 5,2 Å - viz **obr. 9**.



Obr. 9: *Vlastnosti vodíkového můstku. Vodorovná osa je vzdálenost mezi skupinami O=C a NH, svislá představuje hodnoty úhlu θ. Hraniční hodnoty jsou naznačeny nepřerušovanou čarou* (Kabsch & Sander, 1983) *[upraveno].*

Elektrostatická energie determinuje přítomnost vodíkové vazby mezi dvěma rezidui. Je počítána podle Coulombovy rovnice (**Rovnice (1)**) a odpovídá součtu interakcí uhlíku C a kyslíku O karbonylové skupiny s vodíkem H a dusíkem N amidové skupiny. Příslušné parciální náboje pro C, O = (+q_1, - q_1), zatímco pro N, H = (-q_2, + q_2). Hodnota $q1 = \pm 0.42 * 1.60217662e - 19$ a $q2 = \pm 0.2 * 1.60217662e - 19$.

Pro každou amidovou skupinu jsou zvažovány pouze dvě nejstabilněji interagující karbonylové skupiny, což není výslovně vyjádřeno v původním článku (Haghighi et al., 2016).

V případě překrytí dvou stavů na jednom reziduu má nejdůležitější prioritou α-helix, následovaný ß-listem.

U delších struktur mohou chybět některé vodíkové můstky, které by tam v ideálním případě měly být. I v tom případě algoritmus přiřadí reziduu uprostřed pravidelné struktury, které svými vlastnostmi neodpovídá, stav odpovídající této pravidelné struktuře.

Součástí článku je i popis jednotlivých SSE motivů. Odtud vychází takzvaná DSSP klasifikace sekundárních struktur, která je široce používána.

5.3.1 DSSP klasifikace SSE

Podle této klasifikace existují dva základní motivy. Dále existuje několik konzervovaných motivů z nich složených. Prvním ze základních motivů je n-turn. Jedná se o vzor ($i \rightarrow i + n$). Je značen jako stav T, který přiřadíme ke zbytku *i*, pokud existuje vodíková vazba mezi CO(i) a NH(i + n) pro n = 3,4,5. Druhým ze základních motivů je bridge, dále dělený na paralelní a antiparalelní. Tyto základní motivy mají složitější pravidla pro anotaci: Paralelní bridge mezi rezidui *i* a *j* anotujeme, pokud existuje HB mezi (i-1, j)-tým a zároveň (j, i+1)-tým reziduem nebo mezi .(j-1, i)-tým a zároveň mezi (i, j+1)-tým reziduem. tedy:

Parallel_bridge(i,j) =: H_bond(i-1, j) && H_bond(j, i+1) nebo H_bond(j-1, i) && H_bond(i, j+1)

V podobné notaci se pravidlo pro ASS antiparalelního bridge zapíše následovně:

Antipar_bridge(i, j) =: H_bond(i, j) && H_bond(j, i) nebo H_bond(i-1, j+1) && H_bond(j-1, i+1)

Stav *paralelní* je ve výstupu programu značen malými písmeny, *antiparalelní* naopak velkými.

Složené motivy jsou Helix, β-ladder, β-hřeben a β-list.

Helixů je několik typů. V zásadě jde o posloupnosti následných motivů n-turn pro n = 3,4,5 délky alespoň dva. Rozlišujeme:

- 4-helix (i, i + 3) =: 4-turn(i - 1) && 4-turn(i) známý jako α-helix, stav H
    - čili H_bond (i - 1, i +3) && H_bond (i, i + 4).
    - Pozn.: H-bond status zbytků i + 1 a i + 2 nerozhoduje.
- 3-helix (i, i + 2) =: 3-turn(i - 1) && 3-turn(i), což je 3_{10}-helix značený stavem G
- 5-helix (i, i + 5) =: 5-turn(i - 1) && 5-turn(i), běžně označovaný π, stav I

Tato definice nepočítá s hraničními zbytky, které mají počáteční a koncové HB tvořící součást helixu (Andersen & Rost, 2005).

Další složený motiv je ß-ladder. Tvoří jej posloupnost jednoho či více následných motivů "bridge" stejného typu. Co se značení týče, paralelní ladders jsou pojmenovávány malými písmeny, antiparalelní pak velkými. Jako ß-hřeben je označen ß-ladder složený z více než jednoho bridge motivu. Pod pojmem ß-list se myslí složený motiv tvořený posloupností jednoho či více "ladder" motivů, které sdílí AK zbytky.

Posledním ze stavů je Bend (stav S), Tak se označují místa s velkými ohyby (alespoň 70°). Ohyb u i-tého AK zbytku se měří jako úhel mezi *i-3* a *i+3* AK zbytkem, tedy:

Bend(i) =: [angle {(C^{α}(i) - C^{α}(i - 2)), (C^{α}(i + 2) - C^{α}(i))} > 70°]

Kromě samotného stavu přiřazuje DSSP každému reziduu i jeho chiralitu na základě dihedrálních úhlů. Ta je buď pravotočivá, značená jako "+" nebo levotočivá, "-". Pro určení se používá úhel:

α(i) =: dihedrální úhel (C^{α}(i - 1), C^{α}(i), C^{α}(i + 1), C^{α}(i + 2)).

Kritéria chirality jsou následující: "+" se přiřadí, pokud 0° < α < 180° a "-" když -180° < α < 0°. Obecně platí, že helixy jsou převážně pravotočivé (+), kdežto ß-ladders levotočivé (-).

5.4 DEFINE

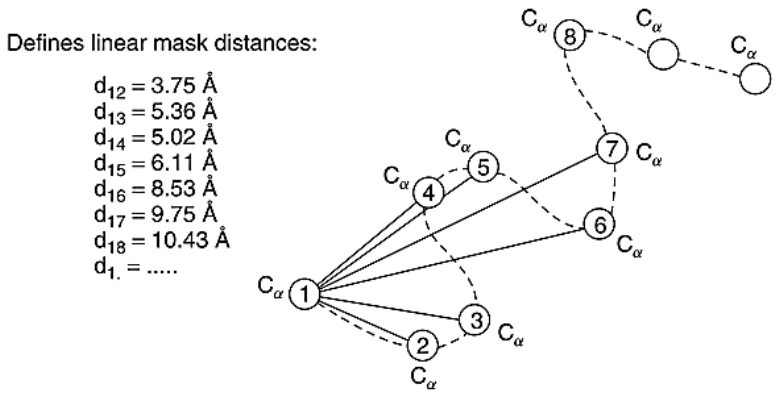
Tento přístup, na rozdíl od DSSP, spoléhá pouze na Cα koordináty a porovnává vzdálenosti Cα atomů se vzdálenostmi v idealizovaných segmentech sekundárních struktur pomocí symetrických *difference distance* matic *(DM)*. Rozdíly mezi ideálními a pozorovanými vzdálenostmi jsou do určité míry akceptovány.

Program určuje málo časté cis-peptidové vazby a následně počáteční a koncové AK zbytky helixů (alfa a jejich 3_{10} konce), ohybů (turns), β-hřebenů a omega smyček, a to postupně v tomto pořadí. Ve většině případů je metoda schopná anotovat 90-95% AK zbytků alespoň jedním typem SSE. Ve druhém kroku anotace se SSE idealizují jako lineární (výpočet směru a umístění os) a kontrolují se jejich běžná zahnutí a zlomy.

Při testování zlomů program vypočítá matici úhlů pro spojované SSE. Pokud je RMS (efektivní hodnota) prvků v matici menší než 25°, je dvojice určena jako jeden zakřivený segment a anotace je odpovídajícím způsobem upravena. Nový SSE je dále reprezentován navzdory zlomu jako lineární segment.

DM jsou indexované od N konce řetězce. Vzdálenost mezi dvěma následnými atomy Cα (i, i + 1) závisí na dihedrálních úhlech, jejichž hodnoty jsou omezeny planaritou peptidové vazby. (i, i + 1) a (i, i + 2) zaujímají úzký rozsah hodnot, protože úhel (i-1, i, i + 1) je v podstatě fixní. Vzdálenosti (i, i + 1) jsou vždy 3,75±0,02Å, zatímco hodnoty (i, i + 2) leží v rozmezí 5,9±0,6Å. Vzdálenosti charakteristické pro SSE se tudíž začínají objevovat až ve třetí poloze (i, i + 3), kde je vzdálenost u helixu rovna 5,0Å (viz **Obr. 10**) a u β-listu přibližně 9,9Å. Pro cis-peptidovou vazbu jsou charakteristické vzdálenosti 2.9±0.2Å.

Anotace helixu probíhá takto: Ideální helix s $C\alpha_i$ na N konci reprezentuje trojúhelník, jehož první řádek (i, i) až (i, i + L-1), obsahuje vzdálenosti k ostatním atomům v helixu. Elementy v dalším řádku budou shodné, ale posunuté doprava a zkrácené o 1 prvek. Vzdálenosti na všech diagonálách jsou tak totožné.



Obr 10: *Lineární maska vzdáleností pro α-helixy. Maska je porovnána se vzdálenostmi v dotazovaném proteinu. Pokud maska odpovídá určitému segmentu, je segment přiřazen jako α-helix. Povolené rozdíly mezi vzdálenostmi v masce a dotazovaném proteinu jsou určeny parametrem tolerance odchylek* (Andersen & Rost, 2005).

Při anotaci helixů je každý Cα považován za možný N koncový atom. Z daného atomu se "rozšiřuje" helix. Cα se v DM pohybuje horizontálně po jednom kroku pro pozice j = i + 1, i + 2 atd. Atom $C\alpha_j$ nechť je výsledný C konec helixu. Všechny vzdálenosti v helixu jsou pak v DM uvedeny v trojúhelníku (i, i), (i, j) a (j, j). Hodnoty v pravém sloupci (i, j), (j, j) představují vzdálenosti $C\alpha_j$ ke všem předchozím atomům v helixu. Získané vzdálenosti jsou srovnávané s odpovídajícími vzdálenostmi v tzv. lineární masce, reprezentující daný SSE. U ostatních SSE probíhá anotace podobně.

Po anotaci α-helixu jsou kontrolovány 3_{10} helikální konce. 3_{10}-helixy jako samostatné SSE tato metoda neuvažuje.

Vzdálenosti v turn elementech jsou velmi podobné jako v helixech. Jakákoli rozumná povolená míra odchylek by způsobila shodu s helixem o délce jedné otočky. Proto je nutné nejdříve anotovat helixy a ohyby uvažovat pouze v oblastech, které nejsou anotovány jako helix.

β-listy jsou strukturně více komplexní. β-hřebeny v paralelních listech jsou mírně kratší než ty v antiparalelních. Minimální povolená délka je 4 rezidua. S maticemi se pracuje stejně jako u helixů.

Délka smyček se pohybuje v rozmezí 6-16 reziduí. Pro každou délku musí mít omega smyčka vzdálenost mezi konci menší, než nějaká předem určená hodnota (10 Å), která musí být zároveň menší než dvě třetiny maximální vzdálenosti uvnitř smyčky.

Smyčky mohou být vytvořeny ze společného počátečního bodu, avšak nesmějí přepsat předchozí SSE anotaci. Jejich anotace je uvedena samostatně.

Kromě anotace SSE je DEFINE následně schopen poskytnout i popis supersekundárních struktur. Počítají se geometrické vztahy mezi lineárními segmenty SSE. Pro určení vztahu je třeba maximálně 6 parametrů (většinou méně).

Určování supersekundární struktury probíhá na 3 úrovních:

1) identifikace interakcí více β-hřebenů nacházejících se v ß-listech pomocí DM
2) charakterizace vzájemné orientace a interakcí mezi β-hřebeny a helixy a mezi dvěma helixy s pomocí metod využívajících kartézské souřadnice
3) identifikace jakékoliv specifikovatelné supersekundární struktury pomocí DM.

Sekundární struktura a jeden aspekt supersekundární struktury jsou reprezentovány “znakovou maticí” (Character Matrix) s formátem odpovídajícím matici vzdáleností. CM umožňuje celkem přesný popis 3D struktury na 2D úrovni. Pro každý SSE je horní trojúhelníková submatice, která představuje všechny vzdálenosti Cα-Cα v tomto SSE, vyplněna znaky identifikujícími tento typ sekundární struktury.

Sousední SSE mohou sdílet krajní AK. Diagonální prvky CM tedy nejsou identifikovány jako SSE. Pokud neexistuje takto sdílené reziduum, koncový AK zbytek SSE je anotován jednoznačně. V opačném případě je AK přiřazena jak předchozímu, tak dalšímu elementu. Součet délky SSE tedy bývá obvykle větší než délka peptidového řetězce.

Kvalita anotace SSE závisí jak na kvalitě dat, tak na zvolené hodnotě parametru tolerance. SSE musí být spojitý, první neshoda s maskou ho tak ukončí. U struktur s malým rozlišením musí být parametr tolerance nastaven poměrně mírně, aby bylo pokrytí sekundární struktury přiměřené.

Používání lineárních masek je dle autorů mnohem objektivnější a nejméně stejně dobré jako vizuální přístup k anotaci. Je také samozřejmě mnohem rychlejší.

Metoda neumí anotovat π-helixy (alespoň se o nich v původním článku nemluví) a turn motivy nekategorizuje do podskupin. Umí však, na rozdíl od DSSP, anotovat nespárované β-hřebeny.

5.5 STRIDE

STRIDE je další z metod, která kombinuje více charakteristik SSE. Zkratka pochází z anglického “STRuctural IDEntification”. Spolu s DSSP patří k nejpoužívanějším metodám.

Podobně jako DSSP pracuje se vzory vodíkových můstků a přiřazuje stejné SSE kategorie. Na rozdíl od něj však používá empirický model vodíkové vazby. Dále bere STRIDE v potaz Φ a Ψ úhly k určení SSE. Cílem je co nejvíce se přiblížit definicím SSE jaké jsou průměrně reprezentovány v ASS krystalografů. Při anotaci musí být brán v úvahu jak vážený vliv vodíkových vazeb tvořících sekundární strukturu tak i torzní úhly kostry.

STRIDE dále bere v potaz statistické faktory pravděpodobností odvozené z empirických vizuálních anotací SSE, extrahovaných z PDB. Pro torzní úhly jsou stanoveny tendence tvořit α-helixy a β-listy (propensity) podle toho, jak blízké jsou jejich hodnoty charakteristickým oblastem v Ramachandranově grafu.

Základní kvantita SSE je vyjádřena jako kombinace dvou motivů: 4-reziduální otočka pro a-helixy a bridges pro β-listy. Jejich kvalita je váženým produktem příslušných energií HB a zmíněných statisticky odvozených tendencí AK zbytků vyskytovat se v α-helixech a β-listech.

Zavedení pouze jedné prahové hodnoty pro tuto kvantitu (pro každý typ vzoru HB) umožňuje přesné ladění parametrů anotace, protože vzory s nepřesnými torzními úhly mohou být stále akceptovány, pokud vytvářejí silné HB vazby. Naopak, relativně slabý vodík může být kompenzován správnou geometrií kostry.

ASS krystalografů, který je stanoven pro stovky dostupných souřadnicových sad, se používá pro ladění prahových hodnot v anotační proceduře. Tento přístup je tedy tzv. “knowledge-based”.

Stejně jako DSSP, STRIDE anotuje α-helix (H), pokud obsahuje alespoň dvě po sobě jdoucí vodíkové vazby i → i +4. Na rozdíl od DSSP jsou helixy prodlouženy, o jeden nebo dva okrajové zbytky, pokud mají přijatelné úhly (Φ / Ψ). Vzory vodíkových vazeb proto mohou být navíc zcela ignorovány, pokud jsou úhly (Φ / Ψ) nepříznivé (což platí i pro krátké helixy).

Metoda nerozlišuje mezi paralelními a antiparalelními listy. Minimální list (E) se skládá ze dvou zbytků v jedné z pěti možných konformací vodíkových vazeb, tj. dvou více než u DSSP. Listy o délce jednoho AK zbytku (β-bridges), jsou značeny jako B pro tři konformace vodíkových vazeb shodných s DSSP definicí a jako b pro zbývající dvě. Dihedrální úhly spoluvytváří ASS listů stejně jako u helixů.

Jak 3_{10}-helixy (G), tak i π-helixy (I) jsou implementovány podle DSSP schématu, ale s empirickým kritériem vodíkové vazby.

Ohyby jsou anotovány na základě (Φ / Ψ) úhlů reziduí i +1 a i +2, jak je popsáno v práci Wilmota a Thorntona. Ti zavedli nový systém turn motivů, protože se ukázalo, že celých 42% pozorovaných otoček nepatřilo (byť těsně) do dosavadních 8 kategorií (Wilmot & Thornton, 1990).

Pro zbývající oblasti, které výše zmíněným kategoriím neodpovídají je v anotaci použit symbol C.

Čistě z definice by měl být STRIDE v anotacích přesnější než DSSP. To dokládá i testování. Dataset byl brán z PDB a představuje reprezentativní vzorek X-ray a NMR struktur. Odebrány byly struktury obsahující pouze C α atomy, krátké struktury (<70 AK), struktury bez autorsky anotovaných SSE, se špatnými anotacemi atd. U použitého datasetu byla anotace každého jedenáctého β-listu a každého dvaatřicátého α-helixu více v souladu s vizuálním ASS než u DSSP anotace.

Nástroj STRIDE našel využití u vizualizace sekundárních struktur nástrojem VMD (Humphrey, Dalke, & Schulten, 1996).

5.6 KAKSI

Algoritmus KAKSI je pojmenovaný podle finského označení pro slovo "dvě". Bere v potaz dvě kritéria, kterými jsou (Φ / Ψ) úhly a vzdálenosti mezi Cα atomy. Specifickým rysem této metody je, že se snaží řešit nepravidelnosti struktur.

Neexistuje žádný standard pravdivosti pro srovnávání metod a jak uznávají i sami autoři článku, nelze říci, že je jejich metodologie zlepšením oproti stávajícím metodám.

Typ	**Jádro**	**Konce**
i → i + 2	5.49(0.20)	5.54(0.25)
i → i + 3	5.30(0.64)	5.36(0.39)
i → i + 4	6.33(0.71)	
i → i + 5	8.72(0.63)	

Tabulka 4: Vzdálenosti v helixech. *Jádro: vzdálenosti uvnitř helixu bez hraničních AK. Konce: vzdálenosti zahrnující nejméně jeden AK zbytek na konci helixu. Střední vzdálenosti, jsou uvedeny v Å s jejich standardními odchylkami v závorkách* (Martin et al., 2005).

Jako “zlatý standard” se používá PDB anotace, přičemž je třeba mít na paměti, že údaje jsou částečně podobné DSSP anotací. Algoritmus využívá posuvná okna, které se v jeho průběhu posouvají po sekvenci. α-helixy jsou přiřazeny jako první, následované detekcí β-listů. Pro α-helix musí být splněno alespoň jedno z kritérií vzdáleností a úhlů. Pro ß-list jsou posouvána po sekvenci dvě okna, protože cílem je přiřadit jen β-hřebeny zapojené do β-listů. Musejí být splněna obě kritéria. Pokud již určité AK zbytky byly anotovány stavem helix, nemohou být určeny jako list.

Algoritmus je specifický kvůli hledání smyček a ohybů v rámci helixů. Detekce ohybů se provádí pouze v jádrech helixů, segmenty koncových reziduí se při výpočtu neberou v úvahu. Jedním ze způsobů, jak detekovat ohyby, je vypočítat vzdálenosti mezi páry (Φ / Ψ) po sobě následujících zbytků j a j+1 v Ramachandranově mapě.

Přiřazené helixy jsou více lineární než ty přiřazené jinými metodami. Algoritmus KAKSI upřednostňuje pravidelnost struktur.

5.6.1 Charakteristiky SSE podle KAKSI

Pro α helix jsou uvažovány čtyři vzdálenosti mezi zbytky i a j v sekvenci, s $j \in [i + 2, i + 5]$. Hodnoty vzdáleností viz **Tabulka 4.** Pro β-list se uvažují tři různé typy vzdáleností, **obr. 9** na str. 10.

Hodnoty Φ a Ψ úhlů pro rezidua podílející se na helixech a listech se počítají z Ramachandranových map, které se rozdělí na 10 čtverců po 10 stupních. Výstupem toho jsou dvě “populační mapy”, jedna je charakteristická pro helixy, druhá pro listy. Pro α-helixy jsou přípustné pouze hodnoty $\Phi < 0°$ a $-90° < \Psi < 60°$.

<u>KAKSI heuristika pro ASS helixů a hřebenů</u>

- Parametry metody jsou: ε_h a ε_b pro definici thresholdů pro Cα vzdálenosti a η_h, σ_b pro thresholdy pro omezení (Φ/Ψ) úhlů.
- Vzdálenostní kritérium pro α-helixy (C1)

    - Všechny vzdálenosti Cα v posuvném okně o délce w1 (nastaveno na 6) musí ležet v intervalu [Mα - ε_H × SDα ;Mα + ε_H × SDα]. Mα a SDα reprezentují střední a směrodatnou odchylku distribucí Cα v a-helixech.
    - Podobný interval funguje u kritéria pro β-řetězce. (C3)
- Úhlové kritérium pro α-helixy (C2)
    - Všechny páry (Φ / Ψ) v posuvném okně o délce w2 (= 4) musí splňovat podmínku (Φ < 0 ° a -90 ° < Ψ <60 °) a alespoň jeden pár musí spadat do zóny populační matice s hustotou > δ_H.
- Úhlové kritérium pro β-listy (C4)
    - Pro každý úhlový pár (Φ / Ψ), který spadá do oblasti Ramachandrana s hustotou > 0 se zvyšuje hodnota čítače o 1. Pokud pro pár (Φ / Ψ) zbytku vprostřed posuvného okna platí -120 ° < Ψ < 50 °, je čítač resetován na nulu. Konečné skóre musí být větší nebo rovno σb.
- Empiricky optimální hodnoty parametrů jsou: ε_H = 1.96, η_H = 2.25, ε_b = 2.58 a σ_b = 5.

5.7 P-SEA

Algoritmus P-SEA (Protein Secondary Element Assignment) je založený výlučně na pozicích Cα, což je obecnější přístup než například DSSP. Použité parametry byly upraveny tak, aby plnily úkol stejně efektivně jako metody založené na analýze kostry. Výhodou je, že metoda řeší zvlášť koncové zbytky SSE, kde často dochází k neshodám.

Metoda je schopna zpracovat dokonce i částečně neúplná data, takže pro první krok k určení struktury stačí pouze znalost stopy Cα atomů. Jako vstup může sloužit libovolný soubor v PDB formátu. Algoritmus přiřazuje pouze 3 stavy: helix (hlavně alfa, ale též 3_{10}- a π-helixy), β-list (paralelní a anti-paralelní β-listy) nebo coil (loops a turns). Čili nerozlišuje mezi jednotlivými typy helixů, listů ani smyček.

Nejprve jsou z kartézských souřadnic Cα uhlíků vypočteny vzdálenosti $d2_i$, $d3_i$ a $d4_i$ mezi (i - 1)-ním zbytkem a (i + 1)-ním a mezi (i + 2) a (i + 3). Dále jsou na základě koordinátů počítány úhly. Úhel T_i je definovaný Cα uhlíkovým tripletem (i - 1, i, i + 1) a dihedrální úhel α_i quadrupletem (i - 1, i, i + 1, i + 2). Přiřazení SSE je následně stanoveno na základě splnění alespoň jednoho z kritérií vzdálenosti ($d2_i$, $d3_i$, $d4_i$) nebo úhlu (T_i, α_i). Použití pouze jednoho typu geometrického kritéria by vedlo k nepřesnému ASS. Kritéria jsou shrnuta v **tabulce 5.**

Parametry byly odvozené z vizuálně anotovaných SSE. Byly zvoleny průměrné hodnoty pozorované v perfektních strukturách. Poté byly parametry vylepšeny přidáním tolerance 10%, aby přiřadily stejné SSE jako krystalograf.

Nejprve jsou přiřazeny helixy, a to k libovolnému pěti nebo více zbytkovému segmentu, ve kterých každá poloha (Cα) má geometrické charakteristiky, které splňují buď vzdálenost ($d3_i$, $d4_i$) nebo (T_i, α_i) úhlová helikální kritéria. Každý segment je pak prodloužen o jeden AK zbytek na každém konci, pokud $d3_i$ nebo ti splňuje helikální kritéria. Poté jsou podobně definovány β-hřebeny podle konkrétních kritérií. Pokud hodnota $d3_i$ odpovídá, je řetězec prodloužen na každém konci o jeden zbytek, podobně jako u helixů. Krátké β-hřebeny (3 AK zbytky) jsou přiřazeny pouze v případě, že jsou součástí β-listů. Loops jsou výchozí stav. Coil je běžně přiřazen k prvnímu a poslednímu zbytku řetězce.

Parametry ASS	Sekundární struktura	
	Helix	β-**list**
Úhel T (°)	89 ± 12	124 ± 14
Dihedrální úhel a (°)	50 ±20	-170 ±45
Vzdálenost d2 (Å)	5.5 ± 0.5	6.7 ± 0.6
Vzdálenost d3 (Å)	5.3 ± 0.5	9.9 ± 0.9
Vzdálenost d4 (Å)	6.4 ± 0.6	12.4 ± 1.1

Tabulka 5: *Parametry algoritmu P-SEA.*

Testování metody proběhlo porovnáváním s algoritmy DSSP, DEFINE, P-CURVE a TCM. Jako dataset sloužila databáze 226 polypeptidových řetězců z PDB (Bernstein et al., 1978), což odpovídá 43 489 AK zbytkům.

Shoda assignmentu s daty z PDB je nižší u beta struktur, což může být dáno mimo jiné jejich kratší průměrnou délkou 7 zbytků oproti >12 u helixů.

Srovnání DEFINE s DSSP nebo P-CURVE ukazuje pouze 75% shodu. P-SEA pří srovnání buď s DSSP nebo P-CURVE dosahuje shody 83% s oběma. Takže P-SEA je zřejmě lepší i když používá pouze souřadnice stopy Cα. DSSP ani P-CURVE nepřekonávají P-SEA. Jejich míra shody s vizuálním průzkumem 3D struktur nebo přiřazení TCM je srovnatelná. Navíc tyto metody potřebují kompletní informaci o struktuře kostry. Ve srovnání s TCM konsenzus metodou je jen minimum krátkých úseků nepřiřazeno.

V současné době se metoda využívá k odvozování databází proteinové struktury pro metody homologního modelování s využitím fragmentů.

5.8 SABLE

Příkladem jedné z novějších metod na anotaci SSE je SABLE. Ten je rozšířením DSSP metody o nový přístup k definici vodíkových můstků. Účelem bylo použít “parameter-free” metodu pro určení vodíkového můstku za účelem větší obecnosti a širšího využití na jakýkoliv systém s vodíkovými vazbami.

Reálný vodíkový můstek je nejsilnější na krátkou vzdálenost a se zvětšující se vzdáleností postupně jeho síla klesá limitně k nule. Pro potřeby anotace je však nutné zavést zjednodušení ve formě stavové funkce nabývající pouze hodnot 0 a 1. Musíme tedy stanovit mez, přičemž její hodnota ovlivní výslednou anotaci. DSSP používá výpočet elektrostatické energie a jako mez je pevně zvolena hodnota −0.5 kcal·mol^{-1}. SABLE za účelem vyhnutí se parametrů používá metodu “Nejsilnějšího akceptoru”, což je v zásadě metoda nejbližšího souseda. Donor vytváří interakci s tím akceptorem, se kterým má nejsilnější elektrostatickou interakci. (Tento postup byl již dříve použit pro vodní prostředí.) Bylo navrženo celkem 12 metod určování vodíkových můstků. Následně byla použita DSSP pravidla pro anotaci. Tři různé proměnné byly použity v kombinaci se třemi způsoby hledání donoru. Jako proměnné se používaly:

1. F_{OH}: Elektrostatická síla mezi O a H = inverzní vzdálenost OH

Náboje O a H jsou pevné => bez parametrů.

2. F_{COHN}: Elektrostatická síla mezi CO a HN se stejnými náboji jako v DSSP
3. E_{COHN}: Elektrostatická energie mezi CO a HN

Metody hledání donoru se lišily v závislosti na tom, jestli má donor vytvořit interakci s jedním, nejsilnějším akceptorem (SA) nebo dvěma (STA) akceptory. Třetí možností (SAB) je vytvořit vazbu s nejsilnějším akceptorem a případně ji rozdělit, pokud druhý nejsilnější akceptor přijímá méně vazeb. (Karbonylové kyslíky mohou přijmout 0-2 HB.)

Tato metoda vychází z předchozí studie o vodě, která ukázala, že donor na kyslík, který již přijímá větší počet vazeb, preferuje rozdělení na sousední kyslík s méně donory.

Nakonec byla ještě přidáno kritérium LE (Local Enviroment), které zamezuje tvorbě příliš dlouhých můstků. Toto kritérium povoluje tzv. *nepřiřazené donory,* což jsou takové, jejichž nejsilnější akceptor je na sousedním reziduu. LE kritérium bylo zkombinováno s F_{OH} a všemi metodami hledání donoru.

Jako metoda s nejlepší shodou s DSSP ASS se ukázala být kombinace F_{OH}+SAB+LE nazvaná SABLE. Testování proběhlo na stejném datasetu jako test KAKSI metody a srovnáním výsledných anotací s DSSP. Výsledné srovnání je vyjádřeno v **Tabulce 6** jako procentuální shoda přítomnosti HB a celkové SSE anotace mezi každou z metod a DSSP.

Metoda hledání donoru	**HB proměnná**			
	E_{COHN}	F_{COHN}	F_{OH}	F_{OH} + LE
SA	80.7 (89.7)	80.4 (88.9)	78.8 (90.7)	88.9 (92.9)
SAB	62.1 (82.2)	62.9 (81.7)	55.1 (85.2)	85.0 (94.6)
STA	42.0 (71.5)	41.5 (63.6)	39.2 (76.4)	76.8 (94.8)

Tabulka 6: Výsledky testování metod. *% shoda v přiřazení HB (anotaci SSE) mezi všemi metodami a DSSP.*

Z těchto měření vyplývá, že LE kritérium rapidně zlepšuje výsledky. Také se ukazuje, že i nejjednodušší metoda SA v kombinaci s F_{OH} stále vykazuje 91% shodu s DSSP.

Co se samotné SABLE metody týče, přiřazuje celkově stejné procento SSE jako DSSP. U obou metod se množství neurčených reziduí pohybuje okolo 3%. Pro α-helixy i β-listy je shoda vysoká, 97%. Mírně nižší shoda byla pozorována pro bends (92%), turns (86%) a bridges (80%). U 3_{10}-helixů je shoda na úrovni 94%, zatímco pro π-helixy je nejhorší - pouze 69%.

DSSP předpovídá o něco více a-helikálních zbytků než SABLE, který místo toho předpovídá více 3_{10}-helixů. Tyto rozdíly jsou většinou na C-koncích helixů, což je typicky jejich nejméně stabilní část, a tudíž nejméně definovaná.

Jednou z nevýhod SABLE je, že hlásí malé množství interakcí s nepříznivou energií, protože používá pouze interakci OH, ale vypočítané energie jsou pro CO-HN. Nepříznivé energie jsou způsobeny příspěvkem atomů C a N. Interakce navíc nejsou povoleny pro postranní řetězce (které mohou být reálným nejlepším akceptorem) a LE kritérium také nemusí být dostatečné, aby zabránilo anotaci dlouhých HB. Nicméně tento přístup je natolik obecný, že se nemusí omezovat pouze na proteiny. S drobnými modifikacemi je použitelný i na určení HB mezi RNA, RNA a proteinem atd. Navíc je teoreticky schopný anotovat i dynamické systémy.

5.9 SECSTR

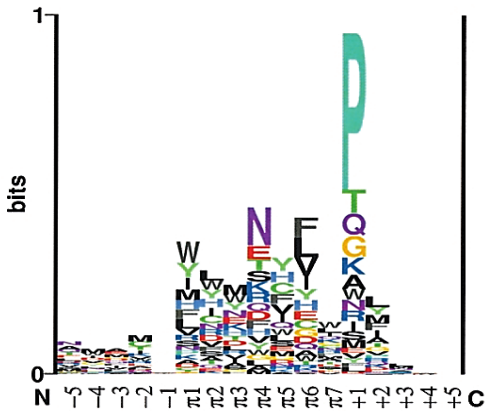
Tato metoda je zaměřena na hledání π-helixů. Jejím hlavním rysem je energetická klasifikace HB. V ASS se pak upřednostňuje vazba s nejvyšší energií.

Použitím zdokonalené definice π-helixů a algoritmu pro jejich hledání se zjistilo, že v neredundantní sadě proteinových struktur s vysokým rozlišením obsahuje π-helixy téměř každý desátý protein. Celkově se množství nalezených π-helixů navýšilo 10x oproti jiným metodám. Ukázalo se, že π-helix má strukturní parametry, které se liší od jeho předpokládaného modelu.

Identifikace π-helixu v algoritmu probíhá pomocí hledání HB vzorů. Definice vodíkové vazby je podobná DSSP. Můstek je přiřazen, pokud má elektrostatickou energii <= 0.5 kcal·mol^{-1}, jako u DSSP. Parciální náboje jsou nicméně přiřazeny podle CNS parametrů (Brünger et al., 1998). Chybějící vodíkové atomy se dopočítávají podle Paulingových pravidel (Pauling & Corey, 1951).

První helikální zbytek je definován jako ten, jehož C=O skupina je zapojena do (i ← i + 5) HB, zatímco poslední zbytek je ten, jehož NH skupina je zapojena do (i - 5 ← i) HB v rámci helixu. Aby se rozlišily π-helixy od π-ohybů, které mají pouze jednu (i ← i + 5) vodíkovou vazbu, minimální π-helix musí mít nejméně dvě (i ← i + 5) HB za sebou.

Důležité je, že pokud existuje více možností tvorby HB, tj. vzory (i←i + 3), (i←i + 4) nebo (i←i + 5), bere se jako platný vzor ten s nejsilnějšími interakcemi. Nedochází tedy k prioritizaci α-helixů jako u DSSP nebo STRIDE.



Obr. 11: Logo graf AK distribuce na 17 pozicích uvnitř a okolo π-helixů. *Výška každého písmena je úměrná jeho frekvenci na této pozici (nejčastější nahoře) a celková výška sloupce je úměrná míře evoluční konzervovanosti. Hydrofobní AK zbytky jsou zbarveny černě, neutrální fialově, kyselé červeně, bazické modře. Prolin je azurový a glycin oranžový. Aromatické AK zbytky jsou bledé (Fodje & Al-Karadaghi, 2002).*

Testovací dataset metody byl vzat z PDB a sestával z 936 proteinových řetězců určených pomocí X-Ray s minimálním rozlišením 2Å a maximální sekvenční identitou 30%. Všechny identifikované π-helixy byly vizuálně prověřené programem SPDBViewer (Guex & Peitsch, 1997). 116 helixů bylo označeno jako potenciální π-helixy, z nichž 104 mohlo být potvrzeno vizuální kontrolou. Několik proteinů mělo více než jeden π-helix. Pro srovnání, DSSP nachází ve stejném datasetu pouze 9 a STRIDE pouze 6 π-helixů.

Z celkových 224 046 AK v datasetu se 728 nacházelo v těchto π-helixech, což odpovídá 0.3%. Délka helixů se pohybovala mezi 7 a 13 AK, nejčastější délka byla 7 AK, což odpovídá 1.5 otočky.

Průměrné hodnoty dihedrálních úhlů u pozorovaných π-helixů jsou (-76 °, -41 °) se standardními odchylkami (σ_Φ, σ_Ψ) = (25, 24). Tyto hodnoty se výrazně liší od navržených modelových hodnot (-57 °, -70 °), ale jsou podobné hodnotám (-77 °, -54 °) pro π-helixy, vzniklé během simulací MD. Struktury s takovýmito úhly jsou mnohem stabilnější. Dále se ukázalo, že hodnoty (Φ / Ψ) jsou pozičně-specifické. Pro pozice π4 a π5 byly střední hodnoty (Φ / Ψ) rovny (-96, -26) a (-97, -51).

Geometrické parametry pozorovaných helixů, jako je počet zbytků na otočku, stoupavost a točivost byly vypočítány pomocí programu HELANAL (Bansal, Kumart, & Velavan, 2000) a velmi odpovídají modelovým hodnotám.

Ve většině případů tvořila C = O skupina prvního AK zbytku π-helixu rozvětvenou vodíkovou vazbu s NH skupinami i + 4 a i + 5, avšak vazba (i ← i + 4) byla slabší.

Pro pouze dva následné HB (i ← i + 5) a (i + 1 ← i + 6) zůstávají tři prostřední skupiny C = O vystaveny solvataci. Podobně pro 3 a 4 následné HB, přičemž u čtyř opakování je pouze 1 C=O skupina volná.

Co se tendencí AK tvořit π-helix týče, zdá se, že v π-helixech jsou obecně výhodné aromatické a velké alifatické AK (Ile, Leu, Tyr, Trp, Phe, His a Asn), zatímco malé AK, jako jsou Ala, Gly a Pro, preferovány nejsou. Velké AK, jako jsou Phe, Trp, Tyr, Ile, Leu a Met, vykazují tendenci nacházet se spíše na koncích helixu. Pro ostatní pozice jsou preferovány polární AK, např. Asn, Glu, Thr a Ser. Asn je typickou středovou AK. Důležitá je extrémní frekvence výskytu Pro na pozici +1. Výskyt prolinu zřejmě ukončuje helix a může mít i stabilizační funkci - možná funguje jako “zátka” středové dutiny. Logo graf π-helixu (Schneider & Stephens, 1990) je na **Obr. 11**.

5.10 P-CURVE

P-Curve je založen na matematické analýze zakřivení proteinů. Pomocí diferenciální geometrie počítá helikální osu na základě superpozice pevného souřadného systému os a lokálních os posloupnosti peptidových rovin. Poskytuje tak úplnou sadu helikálních parametrů a celkovou helikální osu, jedinečnou pro každý protein. Pro optimalizaci těchto parametrů se využívá rozšířená procedura minimalizace nejmenších čtverců. Strukturní nepravidelnosti jsou reprezentovány jako změny orientace následných peptidových rovin a zakřivení celkové helikální osy.

Výhodou tohoto přístupu k anotaci je, že funkce popisující ASS je konstruována tak, že současně zohledňuje polohu všech monomerních jednotek tvořících polymer. Konečný ASS kterékoli z těchto jednotek tak závisí na poloze jeho sousedů. To vede k mnohem koherentnějšímu pohledu na celkovou konformaci, než je tomu u všech čistě lokálních parametrů, jako jsou torzní úhly kostry používané u jiných metod. Zároveň je možné získat detailní informace o umístění jednotlivých peptidů.

Samotný ASS se provádí pomocí porovnávání s motivy, které jsou založeny na standardních hodnotách pro helikální parametry. P-curve ASS se významně liší od ostatních metod, protože používá odlišné parametry (např. poloměru helixu, sklon k ose a twisting).

Používají se následující motivy: pravotočivý a levotočivý alfa-helix, 3_{10}- a π-helix, paralelní a antiparalelní β-listy a některé další struktury. Podobně jako DEFINE může i P-Curve anotovat nespárované β-hřebeny.

Algoritmus P-Curve lze aplikovat v popisech konformačních fragmentů proteinů a hledání takových fragmentů v databázích proteinových struktur. To může být užitečné i k automatickému porovnání podobných struktur.

5.11 ScrewFit

ScrewFit je jedna z méně užívaných metod, proto jeho fungování popíši pouze rámcově. Používá se především pro struktury s nižším rozlišením. Jedním z jejich využití je implementace v nástroji ProMotif, který analyzuje proteinový soubor 3D koordinátů a poskytuje podrobnosti o strukturních motivech proteinu (Hutchinson & Thornton, 2008).

Jako kritérium ASS bere ScrewFit vzdálenosti mezi Cα atomy. Rigorózní matematický popis sekundární struktury lze získat aplikací teorie „screw motions“, ve které je konformace hlavního řetězce popsána ve smyslu lokálních helikálních parametrů.

Metoda používá čtveřice hodnot pro popis superpozice po sobě následujících peptidových rovin v hlavním řetězci. Popisuje sekundární strukturu z hlediska orientačních vzdáleností mezi následnými rovinami a lokálních parametrů helixu. Tyto parametry jsou odvozeny z nejlepší superpozice po sobě jdoucích peptidových rovin. "Nejhorší" shoda se používá k definování orientační vzdálenosti mezi rovinami.

Analýza standardních motivů SSE proteinů patřících do různých tříd foldů ukázala, že všechny běžné motivy jsou dobře rozlišeny orientačními vzdálenostmi a že parametry jako průměr helixu jsou užitečné pro charakterizaci neideálních SSE.

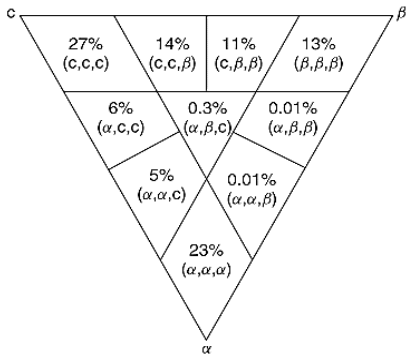
5.12 Další přístupy

Tento výčet metod pochopitelně není úplný. V této práci jsem se snažil zaměřit na více používané metody. VoTAP (Dupuis, Sadoc, & Mornon, 2004) nebo třeba SST (Konagurthu, Lesk, & Allison, 2012) jsou méně používané, množství citací je nižší než u popsaných metod. Konkrétně VoTAP má na WoS pouze 41 citací a SST 12.

SST vyvozuje sekundární strukturu na základě Bayesovské metody - Minimum message length. VoTAP zase anotuje ze seznamu Cα souřadnic a používá 3D Voronoï tessellation.

6 Srovnání algoritmů

Colloc'h a kol. ukázali, že procentuální shoda mezi metodami DSSP, P-CURVE a DEFINE je jen 63%. Tyto nesrovnalosti je přiměly k tomu, aby navrhli metodu počítající ASS na základě trojího konsenzu (TCM) výše zmíněných tří metod (Colloc'h et al., 1993; Labesse et al., 1997). Procentuální shody metod jsou znázorněny na **Obr. 12**. Z něj je patrné, že rozdíl mezi helixem a listem je jasný, protože (α, α, β) a (α, β, β) jsou vzácné (<0,01%).



Obr. 12: Srovnání tří programů pro ASS. *Výskyt tří tříd přiřazení (a-helix, β-list a nepravidelný) určený třemi způsoby anotace: DSSP, P-Curve a DEFINE dávají 10 zobrazených kategorií, pokud se neuvažuje pořadí. Když všechny metody přiřadí α-helix, je to označeno (α, α, α), když dva přiřadí a-helix a jeden nepravidelný, to je označeno (α, α, c)* (Andersen & Rost, 2005).

Jeden ze zajímavých způsobů, jak porovnávat metody anotace je použití benchmarků pro sequence-alignment. Tento přístup použili Zhang a kol., kteří použili tři široce pořizované benchmarky, mezi nimi i PREFAB, který je součástí MUSCLE (Edgar, 2004)**.** Zjistili, že STRIDE a KAKSI dosahují srovnatelné míry úspěšnosti anotace SSE. Jejich úspěšnost je o 1-4% vyšší než u ostatních testovaných metod. Konsensus metod STRIDE, KAKSI, SECSTR a P-SEA, nazvaný SKSP, zlepšuje ASS každém benchmarku o další 1% oproti jednotlivým metodám (W. Zhang, Dunker, & Zhou, 2008).

Procentuální shody v ASS pěti běžně používaných metod na základě benchmarku PREFAB jsou shrnuty v **Tabulce 7**. Z ní je patrné, že jednotlivé metody se navzájem v ASS různě liší. Nejmenší shoda je však těsně pod hranicí 82%. Jednotlivé metody se většinou shodnou na anotaci stavu reziduí nacházejících se uprostřed SSE. Je však často obtížné jednoznačně rozhodnout, které zbytky na okraji segmentů musí být zahrnuty. Tyto hraniční oblasti se označují "twilight zone" (Martin et al. 2005).

%	DSSP	STRIDE	KAKSI	SECSTR	P-SEA
DSSP	100,0	95,4	83,6	94,0	82,1
STRIDE		100,0	85,0	92,1	83,3
KAKSI			100,0	83,2	82,8
SECSTR				100,0	81,9
P-SEA					100,0

Tabulka 7: *porovnání procentuálních shod mezi metodami použitím PREFAB benchmarku.*

7 Závěr

V této práci jsem shrnul vlastnosti SSE. Dále jsem provedl analýzu jednotlivých metod pro ASS. Celkově jsem popsal nejpoužívanější metody a dále jsem vybral další, které sice nejsou tak často citované, nicméně jsou nějak zajímavé. Konkrétně KAKSI metoda je v článku velmi dobře popsána a srovnávána s větším počtem existujících metod. SABLE zase dokazuje, že pomocí jiné definice vodíkových můstků můžeme najít nová uplatnění pro existující metody (hledání interakcí v jiných systémech s HB).

Zajímavé je, že ke stejnému problému lze přistupovat různě, a přesto se alespoň v hrubých rysech metody postavené na jiných základech shodnou na výsledku. Stále však platí, že spolehlivou metodou je manuální anotace krystalografy. Nicméně, vzhledem k časové a finanční náročnosti je jakákoliv automatická metoda anotace s dostatečnou shodou s manuální anotací vítaným a užitečným přínosem.

Vlastní popis a porovnání metod pomohly zjistit, které přístupy existují, jak jsou implementovány a jak moc se jejich výsledky liší. V budoucnu je možné použít některou z metod pro implementaci do MolQL. Jako neslibnější alternativa k DSSP se jeví metoda STRIDE, která je také velmi hojně používána, popřípadě KAKSI. Pokud bychom zůstali u DSSP, pravděpodobně by bylo dobré použít prioritizaci SSE na základě nejvýhodnějších energií, jak je popsáno u SECSTR metody.

Vzhledem k tomu, že naprostá většina struktur v PDB je určena X-ray metodou, která není obvykle schopná určit pozici vodíkových atomů, některá z metod spoléhající pouze na Cα koordináty může být zajímavou alternativou při anotaci SSE.

8 Použitá literatura

Adzhubei, A. A., Sternberg, M. J. E., & Makarov, A. A. (2013). Polyproline-II helix in proteins: Structure and function. *Journal of Molecular Biology*. https://doi.org/10.1016/j.jmb.2013.03.018

Alemán, C., Bianco, A., & Venanzi, M. (2013). *Peptide Materials: From Nanostructures to Applications*. (C. Alemán, A. Bianco, & M. Venanzi, Eds.), *Peptide Materials: From Nanostructures to Applications*. Chichester, UK: John Wiley & Sons. https://doi.org/10.1002/9781118592403

Andersen, C. a F., & Rost, B. (2005). Secondary structure assignment. In *Structural Bioinformatics* (Vol. 44, pp. 341–363). https://doi.org/10.1002/0471721204.ch17

Baker, E. N., & Hubbard, R. E. (1984). Hydrogen bonding in globular proteins. *Progress in Biophysics and Molecular Biology*, *44*(2), 97–179. https://doi.org/10.1016/0079-6107(84)90007-5

Bansal, M., Kumart, S., & Velavan, R. (2000). HELANAL: A Program to Characterize Helix Geometry in Proteins. *Journal of Biomolecular Structure and Dynamics*, *17*(5), 811–819. https://doi.org/10.1080/07391102.2000.10506570

Barlow, D. J., & Thornton, J. M. (1988). Helix geometry in proteins. *Journal of Molecular Biology*, *201*(3), 601–619. https://doi.org/10.1016/0022-2836(88)90641-9

Bawono, P., & Heringa, J. (2014). PRALINE: A versatile multiple sequence alignment toolkit. In *Methods in Molecular Biology* (Vol. 1079, pp. 245–262). https://doi.org/10.1007/978-1-62703-646-7_16

Baybutt, R. B. (1952). The π helix-a hydrogen bonded configuration of the polypeptide chain. *Journal of the American Chemical Society*. https://doi.org/10.1021/ja01142a539

Berman, H. M. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235

Berndt, K. D. (1996). Helices [cryst.bbk.ac.uk]. Retrieved April 17, 2018, from http://www.cryst.bbk.ac.uk/PPS2/course/section8/ss-960531_5.html

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., … Tasumi, M. (1978). The protein data bank: A computer-based archival file for macromolecular structures. *Archives of Biochemistry and Biophysics*, *185*(2), 584–591. https://doi.org/10.1016/0003-9861(78)90204-7

Bowie, J., Luthy, R., & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, *253*(5016), 164–170. https://doi.org/10.1126/science.1853201

Branden, C. I., & Tooze, J. (1999). *Introduction to Protein Structure*. Garland Pub. Retrieved from https://www.ncbi.nlm.nih.gov/nlmcatalog/?term=101503077%5Buid%5D

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., … Warren, G. L. (1998). Crystallography & NMR System: A New Software Suite for Macromolecular Structure Determination. *Acta Crystallographica Section D Biological Crystallography*, *54*(5), 905–921. https://doi.org/10.1107/S0907444998003254

Chothia, C., & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, *5*(4), 823–826. https://doi.org/060 fehlt

Colloc'h, N., Etchebest, C., Thoreau, E., Henrissat, B., & Mornon, J. P. (1993). Comparison of three algorithms for the assignment of secondary structure in proteins: The advantages of a consensus assignment. *Protein Engineering, Design and Selection*, *6*(4), 377–382. https://doi.org/10.1093/protein/6.4.377

Cooley, R. B., Arp, D. J., & Karplus, P. A. (2010). Evolutionary Origin of a Secondary Structure: α-Helices as Cryptic but Widespread Insertional Variations of α-Helices That Enhance Protein Functionality. *Journal of Molecular Biology*, *404*(2), 232–246. https://doi.org/10.1016/j.jmb.2010.09.034

Crick, F. H. C. (1952). Is alpha-keratin a coiled coil? *Nature*, *170*(4334), 882–883. https://doi.org/10.1038/170882b0

Dupuis, F., Sadoc, J. F., & Mornon, J. P. (2004). Protein Secondary Structure Assignment

Through Voronoï Tessellation. *Proteins: Structure, Function and Genetics*, *55*(3), 519–528. https://doi.org/10.1002/prot.10566
Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340
Eisenberg, D. (2003). The discovery of the -helix and -sheet, the principal structural features of proteins. *Proceedings of the National Academy of Sciences*, *100*(20), 11207–11210. https://doi.org/10.1073/pnas.2034522100
Fodje, M. N., & Al-Karadaghi, S. (2002a). Occurrence, conformational features and amino acid propensities for the π-helix. *Protein Engineering, Design and Selection*, *15*(5), 353–358. https://doi.org/10.1093/protein/15.5.353
Fodje, M. N., & Al-Karadaghi, S. (2002b). Occurrence, conformational features and amino acid propensities for the π-helix. *Protein Engineering, Design and Selection*, *15*(5), 353–358. https://doi.org/10.1093/protein/15.5.353
Formaggio, F., Moretto, A., Crisma, M., & Toniolo, C. (2013). Chemistry of Peptide Materials: Synthetic Aspects and 3D Structural Studies. In *Peptide Materials: From Nanostructures to Applications* (pp. 39–63). Wiley-Blackwell. https://doi.org/10.1002/9781118592403.ch2
Frishman, D., & Argos, P. (1995). Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Bioinformatics*, *23*(4), 566–579. https://doi.org/10.1002/prot.340230412
Guex, N., & Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis*, *18*(15), 2714–2723. https://doi.org/10.1002/elps.1150181505
Haghighi, H., Higham, J., & Henchman, R. H. (2016). Parameter-Free Hydrogen-Bond Definition to Classify Protein Secondary Structure. *Journal of Physical Chemistry B*, *120*(33), 8566–8570. https://doi.org/10.1021/acs.jpcb.6b02571
HOŠTÁKOVÁ, N. (2012). *Metody predikce sekundární struktury proteinů*. Brno. Retrieved from http://hdl.handle.net/11012/12442
Hubbard, T. J. P., Ailey, B., Brenner, S. E., Murzin, A. G., & Chothia, C. (1999). SCOP: A structural classification of proteins database. *Nucleic Acids Research*, *27*(1), 254–256. https://doi.org/10.1093/nar/27.1.254
Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, *14*(1), 33–38. https://doi.org/10.1016/0263-7855(96)00018-5
Hutchinson, E. G., & Thornton, J. M. (1993). The greek key motif: Extraction, classification and analysis. *Protein Engineering, Design and Selection*, *6*(3), 233–245. https://doi.org/10.1093/protein/6.3.233
Hutchinson, E. G., & Thornton, J. M. (2008). PROMOTIF-A program to identify and analyze structural motifs in proteins. *Protein Science*, *5*(2), 212–220. https://doi.org/10.1002/pro.5560050204
Ireta, J. (2018). Potential-Energy Surface of Infinite Helical Polypeptides. Retrieved from https://www.researchgate.net/publication/252273054_Potential-Energy_Surface_of_Infinite_Helical_Polypeptides
Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, *22*(12), 2577–2637. https://doi.org/10.1002/bip.360221211
Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., & Phillips, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, *181*(4610), 662–666. https://doi.org/10.1038/181662a0
Kneller, G. R., & Calligari, P. (2006). Efficient characterization of protein secondary structure in terms of screw motions. *Acta Crystallographica Section D: Biological Crystallography*, *62*(3), 302–311. https://doi.org/10.1107/S0907444905042654
Konagurthu, A. S., Lesk, A. M., & Allison, L. (2012). Minimum message length inference of secondary structure from protein coordinate data. *Bioinformatics*, *28*(12), i97-105. https://doi.org/10.1093/bioinformatics/bts223

Labesse, G., Colloc'h, N., Pothier, J., & Mornon, J. P. (1997). P-SEA: a new efficient assignment of secondary structure from C alpha trace of proteins. *Computer Applications in the Biosciences : CABIOS*, *13*(3), 291–295. https://doi.org/10.1093/bioinformatics/13.3.291

Landschulz, W., Johnson, P., & McKnight, S. (1988). The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins. *Science*, *240*(4860), 1759–1764. https://doi.org/10.1126/science.3289117

Law, E. C., Wilman, H. R., Kelm, S., Shi, J., & Deane, C. M. (2016). Examining the Conservation of Kinks in Alpha Helices. *PLoS ONE*, *11*(6), e0157553. https://doi.org/10.1371/journal.pone.0157553

Levitt, M. (1980). Computer Studies Of Protein Molecules. *Protein Folding, R. Jaenicke Editor*, (061/8), 17. https://doi.org/061/8

Martin, J., Letellier, G., Marin, A., Taly, J. F., De Brevern, A. G., & Gibrat, J. F. (2005). Protein secondary structure assignment revisited: A detailed analysis of different assignment methods. *BMC Structural Biology*, *5*, 17. https://doi.org/10.1186/1472-6807-5-17

Mrozek, D., Wieczorek, D., Małysiak-Mrozek, B., & Kozielski, S. (2010). PSS-SQL: Protein secondary structure - Structured query language. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'10* (Vol. 2010, pp. 1073–1076). IEEE. https://doi.org/10.1109/IEMBS.2010.5627303

North, A. C. T. (1992). Protein Architecture: A Practical Approach. *Biochemical Education*, *20*(2), 125–126. https://doi.org/10.1016/0307-4412(92)90141-8

Novotny, M., & Kleywegt, G. J. (2005). A survey of left-handed helices in protein structures. *Journal of Molecular Biology*, *347*(2), 231–241. https://doi.org/10.1016/j.jmb.2005.01.037

Orengo, C., Michie, A., Jones, S., Jones, D., Swindells, M., & Thornton, J. (1997). CATH – a hierarchic classification of protein domain structures. *Structure*, *5*(8), 1093–1109. https://doi.org/10.1016/S0969-2126(97)00260-8

Pal, L., Dasgupta, B., & Chakrabarti, P. (2005). 310-Helix adjoining α-helix and β-strand: Sequence and structural features and their conservation. *Biopolymers*, *78*(3), 147–162. https://doi.org/10.1002/bip.20266

Pauling, L. (1928). The Shared-Electron Chemical Bond. *Proceedings of the National Academy of Sciences*, *14*(4), 359–362. https://doi.org/10.1073/pnas.14.4.359

Pauling, L., & Corey, R. B. (1951). Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds: Two New Pleated Sheets. *Proceedings of the National Academy of Sciences*, *37*(11), 729–740. https://doi.org/10.1073/pnas.37.11.729

Pavone, V., Di Blasio, B., Santini, A., Benedetti, E., Pedone, C., Toniolo, C., & Crisma, M. (1990). The longest, regular polypeptide 3(10) helix at atomic resolution. *J Mol Biol*, *214*(3), 633–635. https://doi.org/0022-2836(90)90279-U [pii]

Petsko, G. A., & Ringe, D. (2004). Protein structure and function (Primers in Biology), *47*(1), A67–A75. https://doi.org/10.1007/BF01952183

Ramachandran, G. N., Ramakrishnan, C., & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, *7*(1), 95–99. https://doi.org/10.1016/S0022-2836(63)80023-6

Richards, F. M., & Kundrot, C. E. (1988). Identification of structural motifs from protein coordinate data: Secondary structure and first-level supersecondary structure. *Proteins: Structure, Function, and Bioinformatics*, *3*(2), 71–84. https://doi.org/10.1002/prot.340030202

Schiffer, M., & Edmundson, A. B. (1967). Use of Helical Wheels to Represent the Structures of Proteins and to Identify Segments with Helical Potential. *Biophysical Journal*, *7*(2), 121–135. https://doi.org/10.1016/S0006-3495(67)86579-2

Schneider, T. D., & Stephens, R. M. (1990). Sequence logos: A new way to display consensus sequences. *Nucleic Acids Research*, *18*(20), 6097–6100. https://doi.org/10.1093/nar/18.20.6097

Simossis, V. A., & Heringa, J. (2005). PRALINE: A multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Research*, *33*(SUPPL. 2), W289–W294. https://doi.org/10.1093/nar/gki390

Sklenar, H., Etchebest, C., & Lavery, R. (1989). Describing protein structure: A general algorithm yielding complete helicoidal parameters and a unique overall axis. *Proteins: Structure, Function, and Bioinformatics*, *6*(1), 46–60. https://doi.org/10.1002/prot.340060105

Suzuki, H., Kolaskar, A. S., Samuel, S. L., Otsuka, J., & Tsugita, A. (1991). A protein secondary structure database (PSS). *Protein Seq Data Anal*, *4*(2), 97–104. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1946337

Taylor, H. S. (1942). Large Molecules Through Atomic Spectacles. *Proceedings of the American Philosophical Society*, *85*(1), 1–12. https://doi.org/10.2307/985121

Toniolo, C., Crisma, M., Formaggio, F., & Peggion, C. (2001). Control of peptide conformation by the Thorpe-Ingold effect (Cα-tetrasubstitution). *Biopolymers - Peptide Science Section*, *60*(6), 396–419. https://doi.org/10.1002/1097-0282(2001)60:6<396::AID-BIP10184>3.0.CO;2-7

Wade, R. C., & Goodford, P. J. (1993). Further Development of Hydrogen Bond Functions for Use in Determining Energetically Favorable Binding Sites on Molecules of Known Structure. 2. Ligand Probe Groups with the Ability To Form More Than Two Hydrogen Bonds. *Journal of Medicinal Chemistry*, *36*(1), 148–156. https://doi.org/10.1021/jm00053a019

Whitford, D. (1961). *Proteins structure and function*. *Deutsche medizinische Wochenschrift (1946)* (Vol. 86). J. Wiley & Sons. Retrieved from https://books.google.fr/books?id=AnodNhuMAdkC

Wilmot, C. M., & Thornton, J. M. (1990). β-turns and their distortions: A proposed new nomenclature. *Protein Engineering, Design and Selection*, *3*(6), 479–493. https://doi.org/10.1093/protein/3.6.479

Zhang, J. (2000). Protein-length distributions for the three domains of life. *Trends in Genetics*, *16*(3), 107–109. https://doi.org/10.1016/S0168-9525(99)01922-8

Zhang, W., Dunker, A. K., & Zhou, Y. (2008). Assessing secondary structure assignment of protein structures by using pairwise sequence-alignment benchmarks. *Proteins: Structure, Function and Genetics*, *71*(1), 61–67. https://doi.org/10.1002/prot.21654

Zhang, Y., & Sagui, C. (2015). Secondary structure assignment for conformationally irregular peptides: Comparison between DSSP, STRIDE and KAKSI. *Journal of Molecular Graphics and Modelling*, *55*, 72–84. https://doi.org/10.1016/j.jmgm.2014.10.005